

# MaReS (Magdeburg Reflective Writing Scoring Rubric for Feedback) – development of a feedback method for reflective writing in health professions education: A pilot study in veterinary medicine

## Abstract

**Aim:** The aim of the study was to develop a scoring rubric that provides valuable feedback to students and to gather evidence for its construct validity.

**Methodology:** The Magdeburg Reflective Writing Feedback and Scoring Rubric (MaReS) was developed in an iterative process following a symposium on reflection by a committee of the “DACH Association for Medical Education (GMA)” in June 2016.

25 essays written by 13 veterinary students were assessed by three independent raters with MaReS and by two raters with the REFLECT rubric in two runs (13 and twelve essays). Validity evidence was gathered referring to the following of Messick’s components of construct validity: content (rubric development), response process (rater manual, rater training, rating time, students’ evaluation), internal structure (inter-rater reliability, IRR), and relationship to other variables (comparison of the rating with the REFLECT rubric and a global rating scale).

**Results:** The analytic rubric comprises twelve items that are rated on three-point rating scales. The authors developed an assignment with guiding questions for students and a rater manual. Results for free marginal kappa of the items of MaReS ranged from -0.08 to 0.77 for the first set of reflective essays and from 0.13 to 0.75 for the second set. Correlations between MaReS and the REFLECT rubric were positive (first run:  $r=0.92$  ( $p<0.001$ ); second run:  $r=0.29$  ( $p=0.37$ )).

**Conclusion:** MaReS might be a useful tool to guide students’ reflective writing and provide structured feedback in health professions education. Using more essays for a rater training and more training cycles are likely to result in higher IRRs.

**Keywords:** reflection, reflective writing, analytic scoring rubric, validity, feedback

Sabine Ramspott<sup>1</sup>

Ulrike Sonntag<sup>2</sup>

Anja Härtl<sup>3</sup>

Stefan Rüttermann<sup>4</sup>

Doris Roller<sup>5</sup>

Marianne Giesler<sup>6</sup>

Linn Hempel<sup>7</sup>

1 Trillium GmbH Medizinischer Fachverlag, Grafrath, Germany

2 Charité – Universitätsmedizin Berlin, Institute of General Practic, Berlin, Germany

3 University Hospital Augsburg, Department of Hygiene and Environmental Medicine, Augsburg, Germany

4 Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz, Germany

5 Ruprecht-Karls-University, Center for Psychosocial Medicine, Institute of Medical Psychology, Heidelberg, Germany

6 Freiburg/Br., Germany

7 Martin-Luther-Universität Halle-Wittenberg, Medical Faculty, Dorothea Erxleben Lernzentrum, Halle/Saale, Germany

# 1. Introduction and aim

## 1.1. Reflection in health professions education

To improve patient care different roles that physicians [16] have to comply with and a set of domains for the veterinary practice [5] have been defined. In order to meet the requirements of these roles or domains healthcare professionals have to acquire many corresponding competences during their studies [<https://www.avma.org/education/accreditation-policies-and-procedures-avma-council-education-coe>], [13], [16], [36], [45]. As demands and knowledge grow, their working environment becomes more and more complex. To provide up-to-date healthcare, physicians and veterinarians consequently have to engage in life-long learning, where the ability to reflect on one's own actions is an integral part [29]. Reflection was already regarded as an essential means of cognition in the Enlightenment (cf. 18<sup>th</sup> century Kant, Fries, among others) and was described as "inner introspection", among other things [17], [25]. Today's literature describes reflection *inter alia* as an "in-depth consideration of events or situations: the people involved, what they experienced, and how they felt about it" [6].

Reflection therefore also supports the process of professional identity formation, a key development process for healthcare professions [9]. Furthermore, reflective practice seems to foster resilience [56] and to reduce stress [33]. These aspects gain more and more importance as the well-being of physicians and veterinarians has become an increasing concern in recent years [19], [31], [32], [55], [60].

The ability to reflect can be developed as part of a learning process [54] and fostered by different activities: Reflective writing has been used widely in medical education [27], [28], [46], [53], [58]. Other activity formats include group discussions [7], [33], online forums [35], digital storytelling [52] and collaborative drawing [34].

The effectiveness of "writing as a method" has been proven [48]. In writing down a perceived situation, people develop the awareness of what had happened and additionally train their expressive skills. This affects how one thinks, feels and acts. People who benefit the most are those who have little opportunity for self-disclosure in everyday life, who find it difficult to recognize and name feelings and who can try out perspectives for their fixed way of thinking while writing [54].

To promote reflection feedback can be a highly relevant and powerful tool, but a systematic review showed that in most of the studies using feedback as part of their intervention to develop reflection the methods and details of the feedback remained unclear [55]. Only one of the studies described the protocol that was used for providing feedback [3].

German speaking countries have many large medical and veterinary faculties with up to 870 first year students

(e.g., LMU Munich, Germany). Reflective writing seems to be an ideal means to engage large numbers of students in reflection. Nevertheless, teaching reflection and providing feedback on reflective writings to a large number of students is expected to be a resource intensive endeavour. Individual written feedback is supposed to be more time consuming than the use of a rubric, where boxes can be ticked. Therefore, scoring rubrics can be a suitable means to simplify feedback. Since the use of structured reflection is relatively new in health professions education in Germany, Austria and Switzerland (DACH region), teachers are usually not very experienced in teaching this competency and providing feedback.

## 1.2. Scoring rubrics

Rubrics are often used to assess performance and reflective writing. There are two different ways of scoring: In holistic scoring the assessors rate the student's essay as a complete unit against a prepared scale, whereas in analytic scoring the assessors break down the essay into its constituent elements, each of which is assigned a proportion of the available mark [22]. Educators usually use holistic scoring for large-scale assessment because it is assumed to be easy to handle, cheap and accurate. Analytical scoring is useful in teaching since the results can help teachers and students to identify students' strengths and learning needs [24]. If the scoring of a reflective writing will be used as a formative assessment, the analytical approach seems more suitable.

Medical educators have developed several holistic [30], [44], [47], [51], [59], [61] and analytic [10], [11], [57] rubrics for the formative and/or summative assessment of reflective essays. Reis et al. developed "The Brown Educational Guide to the Analysis of Narrative" (BEGAN), a guide for crafting written feedback to students' reflective writing [50]. This guide and the analytic rubrics can be used to provide feedback on students' reflective essays. The REFLECT rubric comprises five major criteria (writing spectrum, presence, description of conflict or disorienting dilemma, attending to emotions, analysis and meaning making) and one optional minor criterion (attention to assignment). Four levels of reflection (habitual action, thoughtful action or introspection, reflection, and critical reflection) are assigned to these criteria. If the level critical reflection is achieved the learning outcomes are also to be defined as transformative or confirmatory learning [57]. The rubric of Devlin et al. comprises four dimensions (descriptive, comparative, personal and critical) with prompts (questions) to guide assessment and feedback [11]. Devi's rubric is based on Koole's indicators to describe the process of reflection [29] and Moon's grading system [41]. Raters grade the essays from A to F [10]. So far, no assessment rubric for reflective writing in a medical context could be identified in German.

### 1.3. Guides for reflective writing

There are several guides for reflective writing in medical education: In the appendix to his AMEE Guide, Sandars [51] gives examples for a student information sheet for undergraduate and postgraduate medical students based on Moon's handbook [41], a template for structured reflection to develop a therapeutic relationship of professional practice after Johns [23] with 20 guiding questions and 11 questions to develop deeper reflection. Aronson et al. developed a reflective learning guide that consists of a structured approach to reflection based on the SOAP format with guiding questions and an information sheet about reflection and strategies for successful reflection. Aronson et al. tested their tool using a holistic scoring rubric [4]. We could not identify an analytic scoring rubric accompanied by guiding questions or a guide for students.

### 1.4. Aim

The aim of the study was to develop a German analytic rubric for the formative assessment of reflective essays, that is aligned with guiding questions for students and provides valuable feedback to students. The use of the rubric should be easy and timesaving for teachers when assessing reflective essays. In addition, the authors' goal was to gather validity evidence for the new tool.

## 2. Methods

### 2.1. Scoring rubric development

The committee for Communicative and Social Competencies (KusK) of the "DACH Association for Medical Education (GMA)" (Gesellschaft für Medizinische Ausbildung, GMA) hosted a symposium on reflection in June 2016 to discuss possibilities to incorporate reflection into the teaching of communication and social skills for medical and veterinary students in Germany. The participants of one workshop examined three existing tools to assess reflective essays [44], [50], [57]. Based on these three tools, a first version of a scoring rubric was developed. After the workshop the scoring rubric was refined using Koole's model of common elements describing the process of reflection [29] as a guide for the content selection. Concerning the structure and wording of the scoring rubric, Moskal's recommendations for scoring rubric development were followed [43]. The feedback rubric was developed together with a matching assignment and guiding questions for students and a rater manual in an iterative process. After several revisions by the authors of this paper and after a first consensus was achieved, the rubric, assignment, and guiding questions, the rater manual and an example essay for reflective writing was sent to six external experts, who were engaged in reflective activities at their medical and veterinary faculties in Germany. The expert group comprised doctors, veterinarians, health scientists and psychologists. The feedback

of the external experts on the tool and its elements was discussed and incorporated into the new version of the tool after consensus. The tool was named Magdeburg Reflective Writing Feedback and Scoring Rubric (Magdeburger Reflexionsskala, MaReS) according to its place of origin.

### 2.2. Evidence for construct validity

"Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" [37]. The assessment tool itself cannot be declared valid (or not valid), but more (or less) validity evidence can be collected to support the proposed interpretations of assessment scores. The context in which the validity evidence is collected remains important: assessment data is more – or less – valid for a specific purpose, meaning or interpretation at a certain point of time and for a specific population for which validity evidence was collected [12]. In this study we collected validity evidence for using the rubric as a formative assessment tool of reflective capacity in reflective essays. The essays on veterinary practice were written by veterinary students who took part in an optional subject on reflection in 2017 and were rated by five raters who didn't know the students; each essay was rated by three raters using the newly developed scoring rubric and two using the REFLECT rubric. We gathered validity evidence for four of the six sources of construct validity named by Messick [38]: content, response process, internal structure and relationship to other variables. Figure 1 shows the study design; figure 2 shows the sources of validity evidence selected for this study. In the following section we describe the methods that we used to obtain evidence for the different sources of construct validity in a chronological order. Therefore, the source "response process" is split up in this section.

#### 2.2.1. Content – feedback rubric development

The way in which we developed the instrument contributes to the evidence for content validity (see 2.1.).

#### 2.2.2. Response process I – use of rater manual and rater training

All five raters attended the two-day KusK-Workshop on reflection in June 2016. After the development of MaReS, raters (us, lh, srü, dr, ah) were provided with the final version of the rubric, assignment, and guiding questions and the rater manual. They rated an example essay individually. Ratings were compared and divergent ratings were discussed with all raters in a telephone conference. Subsequently all raters agreed on a common approach for the items. The rater manual was adjusted accordingly. The same approach was followed for the REFLECT rubric [58] that we employed to evaluate the relationship to

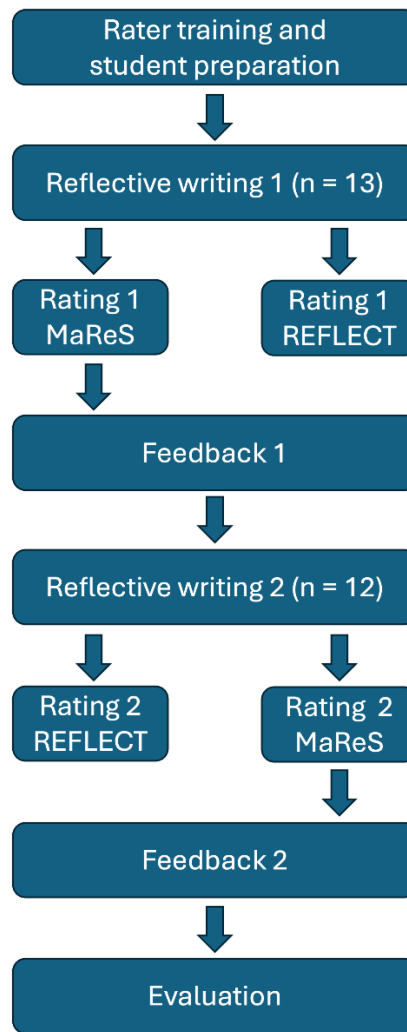


Figure 1: Study design MaReS

<b>Content</b>	<ul style="list-style-type: none"> <li>• Use of Koole’s model of common elements of reflection</li> <li>• Use of Moskal’s recommendations for scoring rubric development</li> <li>• Discussion with experts</li> </ul>
<b>Response Process</b>	<ul style="list-style-type: none"> <li>• Development and use of a rater manual, rater training</li> <li>• Rating time</li> <li>• Student evaluation</li> </ul>
<b>Internal Structure</b>	<ul style="list-style-type: none"> <li>• Inter-rater reliability</li> </ul>
<b>Relationship to other variables</b>	<ul style="list-style-type: none"> <li>• Comparison to another instrument (REFLECT and a global rating scale)</li> </ul>

Figure 2: Sources of validity evidence for MaReS

other variables (see 2.3.). Since there was no rater manual at that time, adjusting the manual did not apply. We used the English version of REFLECT and the four steps for its application [57].

### 2.2.3. Internal structure – inter-rater reliability and relationship to other variables – comparison to another scoring rubric

The Ethics Committee of the LMU Munich granted ethical approval for the study (project number 17-065). Thirteen students of an elective course about reflection on veterinary practice each wrote one reflective essay after six hours of instruction about reflection in general and different reflective activities, including reflective writing based on MaReS. Students were made familiar with the assessment tool. They discussed different examples of reflective writing and the scoring with MaReS. The assignment was to *write a reflective essay about a concrete situation. The students had to choose a situation connected to their studies or occupation in which they have felt challenged during the interaction with a patient, a patient owner, a fellow student or a teacher.* The 13 essays were assessed by three raters using MaReS (ah, dr, lh) and by two raters using REFLECT (us, srü). One of the MaReS ratings was reported back to the students. Subsequently twelve of the students wrote another reflective essay that was rated with MaReS by three raters again (us, srü, lh) and with REFLECT by two raters (ah, dr). One of the MaReS ratings was again reported back to the students. All raters filled in a five-point scale overall rating for every essay. We used the overall and REFLECT rating for study purposes only and not reported back to the students.

### 2.2.4. Response process II – rating time and rater notes

Raters also reported the time they needed for their rating, and, if necessary, made notes on important aspects of the essay, reported difficult items and possible new anchor examples for the rater manual.

### 2.2.5. Response process III – student evaluation

After receiving two feedback sheets on their essays with MaReS, eight students filled in a questionnaire. It comprised five general questions on reflection and the elective subject, six questions on the assignment and guiding questions, four questions on the feedback they received and five questions on their reflective ability and views on reflective writing.

## 2.3. Statistical methods

The rating time for the essays was analysed by descriptive statistics using mean values and standard deviation (SD). Inter-rater reliability (IRR) was determined using free marginal kappa [49]. Correlation between overall MaReS

score (sum of the overall rating for MaReS of 3 raters), overall REFLECT score (sum of the overall rating for REFLECT of 2 raters) and global rating (sum of the corresponding global rating of 3 raters for MaReS and of 2 raters for REFLECT) was determined using Spearman's Rank correlation coefficient.

Students' evaluations and raters' feedback were analysed using descriptive statistics (absolute numbers) for questions with answers on a five-point scale and using content analysis according to Mayring for open questions.

## 3. Results

### 3.1. Scoring rubric development

We developed a writing assignment with nine guiding questions and a rubric with twelve items that were assessed with a three-point scale (see attachment 1). The rater manual consisted of a short description of how to rate the essay and descriptors and anchors for all items and consecutive scales.

### 3.2. Evidence for construct validity

#### 3.2.1. Content – MaReS development

See 3.1.

#### 3.2.2. Response process – rating time

The number of characters (including spaces) of the reflective essays ranged from 2,765 to 8,488 with a mean of 5,359 characters (SD 1,660). The mean time that was needed to rate an essay with MaReS was 13.9 minutes with a standard deviation of 10.9. The mean rating time, i.e., time to read and rate the essay differed greatly between raters, ranging from a mean of 6.7 minutes (SD 2.5) to 28.5 minutes (SD 13.1). The mean duration of the REFLECT ratings was 7.8 minutes with a standard deviation of 5.2. Mean rating times for the individual raters ranged from 4.5 minutes (SD 1.2) to 14.5 (SD 6.9).

#### 3.2.3. Internal structure – inter-rater reliability

Results for free marginal kappa of the items of MaReS ranged from -0.08 to 0.77 for the first set of reflective essays and from 0.13 to 0.75 for the second set (see table 1). Results for free marginal kappa of the individual criteria of REFLECT ranged from -0.26 to 0.31 for the first set of reflective essays and from -0.11 to 0.38 for the second set (see table 2). Free marginal kappa for the global rating of all five raters was 0.16 for the first set of essays and 0.22 for the second set of essays.

**Table 1: Inter-rater reliability for MaReS and REFLECT for the different raters**

Item/ Criterion	Free marginal kappa			
	MaReS		REFLECT	
	AH, DR, LH	US, SRÜ, LH	US, SRÜ	AH, DR
1	<b><u>0.77</u></b>	<b><u>0.75</u></b>	0.28	0.22
2	<b><u>0.77</u></b>	<b>0.67</b>	0.08	-0.11
3	<b>0.62</b>	<b><u>0.75</u></b>	0.18	0.33
4	<b>0.65</b>	0.58	-0.26	0.00
5	0.39	0.25	0.28	0.22
6	0.08	<b><u>0.75</u></b>	-0.03	0.22
7	-0.08	0.13		
8	0.19	0.13		
9	<b>0.62</b>	0.42		
10	0.04	0.25		
11	0.35	0.33		
12	0.08	0.25		
		Transformative vs. Confirmatory Learning	0.31	0.38
		Overall Rating	0.18	0.22

Values between .60 and .74 are considered as “good” and marked with bold numbers.

Values between .75 and 1 are considered as “excellent” and marked with bold and underlined numbers.

### 3.2.4. Relationship to other variables – correlation between ratings

For essays 1 to 13 Spearman’s rank correlation coefficient for MaReS and the raters corresponding overall rating was  $r=0.43$  ( $p=0.14$ ),  $r=0.92$  ( $p<0.001$ ) for MaReS and REFLECT and  $r=0.83$  ( $p<0.001$ ) for REFLECT and global rating.

For essays 14 to 25 Spearman’s Rank correlation coefficient for MaReS and the raters’ corresponding overall rating was  $r=0.75$  ( $p=0.005$ ),  $r=0.29$  ( $p=0.37$ ) for MaReS and REFLECT and  $r=0.87$  ( $p<0.001$ ) for REFLECT and the raters’ corresponding global rating.

### 3.2.5. Response process – students’ evaluations

Eight students participated in the evaluation. The results of the closed questions are shown in table 2. All students rated their own writing competence positive and did understand the assignment. The difficulty of the assignment (“writing a reflection report on a self-selected challenging situation”) was rated differently by the students. For the first reflection report, most students rated their own selection of situations as “perfectly suitable” or “suitable”. For the second reflection report, the selection was rated as “partially suitable” by most students.

All eight students rated the guiding questions as “very helpful” or “helpful”. Knowing that they will receive feedback on their reflection report was mostly rated “conducive” or “rather conducive”. When writing the second reflection report, half of the students felt “slightly more confident”, while the other half felt “slightly less confident”.

ent”. Students rated their own ability to reflect mainly good (four students). Four students stated that they found the second reflection more difficult because they had difficulty finding a suitable situation to reflect on.

The structured feedback was rated positively by five students. They wrote that the feedback was non-judgmental, constructive, and helpful. At the same time one student criticized the fact that the feedback only related to the process of reflection and not to the situation. Thinking about their own ability to reflect, five students stated that they thought they are very self-critical and that this makes reflection difficult. Other responses where: not being able to think of a solution immediately; need for more opinions from others; more time to reflect.

The content analysis revealed the following: In their free text responses, several students cited stress management/reduction as an expected personal benefit of reflection. Additionally personal development in various fields, improved understanding of and communication with others (including in private), benefits for future action in general and in learning situations were mentioned. Additionally, all free-text responses to three questions about the impact of the writing of and the feedback on reflection reports can be found in table 3.

## 4. Discussion

### 4.1. Scoring rubric development

Practicality was our main priority for MaReS. The aim was to develop a scoring rubric that was easy to use and timesaving for the teachers. The poorer inter-rater reliab-

Table 2: Answers to closed questions of the students' evaluation

Question		Characteristic (Number of answers)				
How would you rate your "writing competence," independent of the reflection reports?		very good (1)	rather good (4)	quite ok (3)	not so good (0)	not good (0)
The assignment for the written reflection was concrete and understandable for me.		fully (6)	mostly (2)	partially (0)	not really (0)	not at all (0)
How would you rate the "difficulty" of the assignment - that is, writing the reflection report on a challenging situation of your choice - for yourself?		very easy (1)	rather easy (2)	partially (4)	rather difficult (1)	very difficult (0)
In hindsight, how do you now evaluate your choice of situations for the reflection reports?	Essay 1	perfectly suitable (4)	suitable (2)	partially suitable (2)	not so suitable (0)	not suitable (0)
	Essay 2	perfectly suitable (0)	suitable (1)	partially suitable (5)	not so suitable (2)	not suitable (0)
How helpful were the guiding questions for writing the reflective essays for you?		very helpful (7)	helpful (1)	partially helpful (0)	not so helpful (0)	not helpful at all (0)
Were there moments while writing the report when you were "just stuck," writing was very difficult, or/and you were hesitant to write what you were thinking/feeling?		yes, again and again (0)	yes, sometimes (3)	only very rarely (0)	no, never (5)	-
To what extent has knowing that you will receive feedback on your reflection report helped or hindered your reflection writing?		conducive (4)	rather conducive (3)	partially (1)	rather obstructive (0)	obstructive (0)
Did you feel more confident or not confident when writing the second reflection report than when writing the first reflection report?		significantly more confident (0)	slightly more confident (4)	unchanged (0)	slightly less confident (4)	less confident (0)
How would you rate your own ability to reflect?		very good (0)	rather good (4)	quite ok (3)	not so good (1)	not good (0)

ility for REFLECT could be an indication that MaReS is easier to use. Mean rating times for MaReS were longer than for REFLECT. This probably reflects the fact that MaReS has a greater number of items ( $n=12$ ) than REFLECT ( $n=6$ ). In our opinion less than 15 minutes to read and rate a reflective essay (mean rating time) are still to be considered reasonable.

## 4.2. Evidence for construct validity

In this study, we gathered validity evidence referring to the following components of construct validity: content, response process, internal structure and relationship to other variables [1], [12], [38]. MaReS might be a suitable rubric for teaching reflection and providing feedback to students written reflections. The combination of guiding questions for written reflection and the use of MaReS as a time-saving way to provide structured feedback was successfully piloted.

### 4.2.1. Content

Validity evidence based on test content derives from an analysis of the relationship of a test content and the construct it is intended to measure [1]. The content of MaReS (including guiding questions, rubric and rater manual) was selected carefully based on a widely accepted model for the concept of the process of reflection [29]. We also considered the content of three existing tools to assess reflective essays [44], [50], [57]. More validity evidence was obtained by consulting six external experts to judge the relationship of parts of the test and the construct [1]. We also incorporated their feedback in the final version of MaReS.

### 4.2.2. Response process

Validity evidence based on the response process can include the response process of the raters as well as the response process of the students taking the assignment [1].

To support the assessment process of the raters, we developed a rater manual that comprised a short description

**Table 3: Free-text responses on the impact of writing and feedback on reflection reports**

<b>To what extent did writing the reflective essay and receiving feedback assist you in improving your own self-awareness regarding the situations described and perhaps beyond?</b>	
Student 1	You think more about what happened and improvements.
Student 2	Through this, I processed the situations and learned to reflect.
Student 3	Hardly, actually only thoughts put into words
Student 4	Longer and more intense thought
Student 5	-
Student 6	It helped me see the situation a little more clearly. However, conversations are more important to me as a way of reflection.
Student 7	strongly
Student 8	Through writing, I felt compelled to address a problem that was weighing heavily on me, so I also dared to be aware of myself at all.
<b>To what extent did the writing of the reflective essays and the feedback help you in terms of your view of how to deal with your own emotions and how to concretize realistic plans of action for similar situations and perhaps beyond?</b>	
Student 1	I now think more about conversations with others and how others feel.
Student 2	It has opened my eyes; (since I also had to look at the perspective of the others). I must not be so sensitive!
Student 3	-
Student 4	Very!
Student 5	-
Student 6	-
Student 7	Strongly -> have become more aware of how to deal with such situations
Student 8	see above (Through writing, I felt compelled to address a problem that was weighing heavily on me, so I also dared to be aware of myself at all.)
<b>Overall, what was the most relevant learning outcome for you, or what insight(s) were you really able to gain from reflective writing and/or feedback?</b>	
Student 1	That I can avoid conflicts through reflection
Student 2	see above (It has opened my eyes; (since I also had to look at the perspective of the others). I must not be so sensitive!)
Student 3	Stress Management
Student 4	One takes care to behave differently in a future situation.
Student 5	-
Student 6	To find forecasts for actions in the future.
Student 7	To reflect -> active Acknowledge mistakes -> do not dramatize
Student 8	I never realized before how important it is to take external sources into account.

of how to rate the essay as well as descriptors and anchor examples for all items and consecutive scales. Rater training included the two-day workshop on reflection, rating an exemplary essay and a discussion of divergent results of the rating. At the end of the discussion agreement on how to rate the essays was achieved. Even though “interrater reliability is enhanced by training data collectors, providing them with a guide for recording their observations” [14], and also that in this study we took care to ensure that all raters understood how essays should be rated, we found a poor inter-rater reliability for most of the items (see 4.2.3.). This might indicate that rater training might not have been sufficient. During the response process data about the rating time was collected. This does not add to the validity evidence; it is meant to help educators in the health professions who would like to incorporate reflective writing into their

curricula to assess the feasibility of the endeavour and plan resources. Validity evidence based on response process of the students included student format familiarity [12]. Before writing their essays, students were handed out the guiding questions and the scoring rubric. The meaning of every item and the corresponding scores was explained. Students were given two example essays and discussed the ratings in class. Another source of validity evidence is understandable and accurate descriptions and interpretations of the scores for students [12]. In their evaluations all eight students stated that the assignment was “fully” or “mostly understandable” and the guiding questions were rated as “very helpful” or “helpful”.



### 4.2.3. Internal structure

According to Fleiss an inter-rater reliability (IRR) of 0.6 and above can be considered as excellent and an IRR above 0.4 and lower than 0.6 as fair to good. IRRs lower than .4 are considered as poor [15]. This means that IRR for MaReS in our study are only acceptable for items 1 to 4 and item 9. Looking at the wording of the items, the content that is assessed in items 1 to 4 and 9 (general comprehensibility, reference to the assignment, description of situation, description of own emotions, selection of external sources) seems to be more concrete and thus easier to rate than that of the other items. Items 5 to 8 (explanation of own emotions, describing the perspective of the counterpart, relating the perspective of the counterpart to own perspective, influence of previous experiences and reflections) and items 10 to 12 (assessment of the situation, action strategy, expectations regarding the use of the future action strategy/-ies) are integral but more difficult parts of reflection. Comprehensibility on the side of the raters seems to be more subjective in these items than for the items with an acceptable IRR. For example, on the one hand the explanation of one's own emotions can be difficult, and students might feel the matter is too personal to share in a reflective essay. On the other hand, whether an explanation is comprehensible for a rater might depend on the rater's personal experiences. The same applies for change of perspective. Items 10 to 12 (assessment of the situation, action plan) can be difficult for raters, because there might be an assessment of the situation by the student that is comprehensible or an action strategy that is concrete, but the rater feels that it is not sufficient to handle a similar situation better in the future. This might cause some of the raters to give lower scores. During the assessment of the essays the raters took notes on items that were difficult to rate. The analysis of these notes is not part of this study but will help to refine the rater manual and address important aspects in future rater training.

Item 7 (relating the perspective of the counterpart to own perspective) and item 8 (influence of previous experiences and reflection) also received comparatively low ratings. This might indicate that these items are too difficult. In the case of item 7 students might need more explanation, examples and training. A reason for low scores for item 8 might be that the students in our study did not have much clinical experience and it was difficult for them to draw on previous experiences or reflections.

All IRRs that we found for the REFLECT rubric must be considered as poor. While three studies have found high IRRs for REFLECT [39], [40], [57], another study was also not able to replicate these high IRRs. The authors of the latter study state that the difference in the findings could originate from the context in which validity evidence is collected (e.g. different institution and study population) [18]. Since in the case of our study, German native speakers applied an English tool, language problems should also be considered. It is also likely that rater training for REFLECT was not sufficient. There is a short

description of the application of REFLECT in the original paper [57], but the raters did not have access to a rater manual. Our rater training for REFLECT consisted of the rating and discussion of only one example essay. Miller-Kuhlman et al. applied a sounder approach for their rater training: In a training cycle the raters compared scores and discussed discrepancies for several sample essays, then rated more samples until an IRR of at least 0.8 was achieved before collecting the data for their study. This training required six hours for REFLECT [39].

### 4.2.4. Relationship to other variables

Comparing a newly developed test to a test hypothesized to measure the same construct, is an important source of construct validity [1]. The authors found a positive correlation for the comparison of the MaReS scores with a global rating of the essays and with the REFLECT rubric, which is used to assess students' reflective levels and is meant to provide individualized written feedback to guide reflective capacity promotion [57]. For the first set of essays, we found strong and significant positive correlations [8] for MaReS and REFLECT as well as for REFLECT and the global rating. For the second set we found strong and significant positive correlations for MaReS and the global rating as well as for REFLECT and the global rating. We found near moderate and moderate positive (defined by Cohen as an  $r$  of .3 [8]) correlation for MaReS/global rating in the first set of essays and MaReS/REFLECT in the second set of essays, which were both not significant probably due to the small sample size. The fact that different correlations were found in the first and second rounds could have been caused by a change of raters. Nevertheless, the results suggest that MaReS measures the same construct as the REFLECT rubric [1], namely the students' reflective level.

### 4.2.5. Evaluation

This study examines data for level 1 (reaction) of the New World Kirkpatrick Model [26]. When looking at the results one must have in mind that the number of students that filled in the evaluation for MaReS was very small ( $n=8$ ), and conclusions must be drawn with caution.

Some students share very personal stories in their reflective essays, and care must be taken that feedback on the essays will not be judgmental [21] and hurt students' feelings. The content of reflection is subjective, which is why Koole et al. suggest that assessment should focus on generic process skills [29]. Therefore, most of the rubrics that are used for assessing reflective essays (including MaReS) focus on the process of reflection rather than on the handling of a situation. Nevertheless, some students seem to feel the need for a comment on the situation, as one student mentioned that the feedback had no impact on him/her because it was only evaluating the structure of the text and not the situation itself. Teachers should also be aware that essays can contain very personal content from the students. For example,

descriptions of emotional burdens, illnesses or traumatic experiences could be included in the reports. Therefore, teachers should think in advance about where support may be available and how they will react in such cases. In effective feedback, information about previous performance is used to promote positive and desirable development [2]. Students found the feedback that they received with MaReS constructive, helpful, and non-judgmental. Gaining evidence on level 2 of the New World Kirkpatrick Model [26] will be challenging for MaReS, because reflective capacity is context dependent. Moniz et al. infer from their study about the use of reflective writing for student assessment that drawing meaningful conclusions about reflective capacity requires approximately 14 writing samples per student, which are each assessed by four or five raters [40]. This conclusion raises questions about the feasibility of the summative assessment of reflective writing. Results of our study also lead in the same direction: Even though the feedback instrument was mostly rated positive, half of the students felt slightly less confident when writing their second essay – and their free responses indicated that this feeling was related to the situation they chose for the second reflection (e.g., problems finding a topic for the second reflection).

## 5. Conclusion

The Magdeburg Reflective Writing Feedback and Scoring Rubric (MaReS) can be used as a tool to guide students' reflective writing and provide structured feedback in health professions education. In this study IRR was low for seven of the twelve items. We theorize that the rater training – consisting of the rating and discussion of one exemplary essay – was not sufficient. Using more essays for a rater training and more training cycles are likely to result in higher IRRs. A mean rating time of 13.9 minutes seems feasible and might be shorter, when raters gain more experience. If educators would like to incorporate reflective writing and its assessment into a curriculum, sufficient time for rater training must be allocated when planning resources. Caution must be taken, when summative assessment is used for reflective writing regardless of the tool that is used, because there is low predictability from one essay to the next.

The small number of students that provided feedback on MaReS considered the instrument as comprehensible and helpful. More studies with a greater number of students will be needed to support these findings. Gathering evidence for Kirkpatrick Levels of two and higher will be challenging because of the context specificity of reflective capacity.

We recommend MaReS as a tool for teaching and formatively assessing written reflections of students in health professions education, for example on clinical experiences during a practical year, clinical rotations or block training, but also on learning experiences in general.

## Authors' ORCIDiDs

- Anja Härtl: [0009-0008-0818-6213]
- Stefan Rüttermann: [0000-0002-2293-8089]
- Linn Hempel: [0009-0009-5421-2029]

## Acknowledgements

We would like to thank the members of the Committee for Communicative and Social Competencies (KusK) and the experts for their great cooperation and support.

## Competing interests

The authors declare that they have no competing interests.

## Attachments

Available from <https://doi.org/10.3205/zma001752>

1. Attachment\_1.pdf (239 KB)  
Magdeburg reflective writing feedback and scoring rubric

## References

1. American Educational Research Association; American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999.
2. Archer JC. State of the science in health professional education: effective feedback. *Med Educ.* 2010;44(1):101-108. DOI: 10.1111/j.1365-2923.2009.03546.x
3. Aronson L, Niehaus B, Hill-Sakurai L, Lai C, O'Sullivan PS. A comparison of two methods of teaching reflective ability in Year 3 medical students. *Med Educ.* 2012;46(8):807-814. DOI: 10.1111/j.1365-2923.2012.04299.x
4. Aronson L, Niehaus B, Lindow J, Robertson PA, O'Sullivan PS. Development and pilot testing of a reflective learning guide for medical education. *Med Teach.* 2011;33(10):e515-e521. DOI: 10.3109/0142159X.2011.599894
5. Bok HG, Jaarsma DA, Teunissen PW, van der Vleuten CP, van Beukelen P. Development and validation of a competency framework for veterinarians. *J Vet Med Educ.* 2011;38(3):262-269. DOI: 10.3138/jvme.38.3.262
6. Bolton G. *Reflective Practice. Writing Professional Development.* Third Edition. London: SAGE; 2010.
7. Chu SY, Lin CW, Lin MJ, Wen CC. Psychosocial issues discovered through reflective group dialogue between medical students. *BMC Med Educ.* 2018;18(1):12. DOI: 10.1186/s12909-017-1114-x
8. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: L. Erlbaum Associates; 1988.
9. Cruess SR, Cruess RL, Steinert Y. Supporting the development of a professional identity: General principles. *Med Teach.* 2019;41(6):641-649. DOI: 10.1080/0142159X.2018.1536260

10. Devi V, Abraham RR, Kamath U. Teaching and Assessing Reflecting Skills among Undergraduate Medical Students Experiencing Research. *J Clin Diagn Res.* 2017;11(1):jc01-jc05. DOI: 10.7860/JCDR/2017/20186.9142
11. Devlin MJ, Mutnick A, Balmer D, Richards BF. Clerkship-based reflective writing: a rubric for feedback. *Med Educ.* 2010;44(11):1143-1144. DOI: 10.1111/j.1365-2923.2010.03815.x
12. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-837. DOI: 10.1046/j.1365-2923.2003.01594.x
13. European Association of Establishments for Veterinary Education (EAEVE); Federation of Veterinarians of Europe (FVE). European System of Evaluation of Veterinary Training (ESEVT). Manual of Standard of Operation Procedure. Wien: EAEVE; 2019. Zugänglich unter/available from: [https://www.eaeve.org/fileadmin/downloads/SOP/ESEVT\\_SOP\\_2019\\_adopted\\_by\\_the\\_32nd\\_GA\\_in\\_Zagreb\\_on\\_30\\_May\\_2019.pdf](https://www.eaeve.org/fileadmin/downloads/SOP/ESEVT_SOP_2019_adopted_by_the_32nd_GA_in_Zagreb_on_30_May_2019.pdf)
14. Fink A. Survey Research Methods. In: Peterson P, Baker E, McGaw B, editors. *International Encyclopedia of Education*. Third Edition. Amsterdam: Elsevier Science; 2010. p.152-160. DOI: 10.1016/B978-0-08-044894-7.00296-7
15. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley-Interscience; 1981.
16. Frank J, Snell L, Sherbino J. *CanMEDS 2015 Physician Competency Framework*. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.
17. Fries JF. *Neue oder anthropologische Kritik der Vernunft*. Bd. 2. Zweite Auflage. 1831.
18. Grierson L, Winemaker S, Taniguchi A, Howard M, Marshall D, Zazulak J. The reliability characteristics of the REFLECT rubric for assessing reflective capacity through expressive writing assignments: A replication study. *Perspect Med Educ.* 2020;9(5):281-285. DOI: 10.1007/s40037-020-00611-2
19. Haramati A, Cotton S, Padmore JS, Wald HS, Weissinger PA. Strategies to promote resilience, empathy and well-being in the health professions: Insights from the 2015 CENTILE Conference. *Med Teach.* 2017;39(2):118-119. DOI: 10.1080/0142159X.2017.1279278
20. Heimes S. Warum Schreiben hilft: Die Wirksamkeitsnachweise zur Poesietherapie. Göttingen: Vandenhoeck&Ruprecht GmbH&Co.KG; 2012. DOI: 10.13109/9783666401619
21. Hewson MG, Little ML. Giving Feedback in Medical Education: verification of recommended techniques. *J Gen Intern Med.* 1998;13(2):111-116. DOI: 10.1046/j.1525-1497.1998.00027.x
22. Hunter DM, Jones RM, Randhawa BS. The use of holistic versus analytic scoring for large-scale assessment of writing. *Can J Program Eval.* 1996;11(2):61. DOI: 10.3138/cjpe.11.00
23. Johns C. Framing learning through reflection within Carper's fundamental ways of knowing in nursing. *J Adv Nurs.* 1995;22(2):226-234. DOI: 10.1046/j.1365-2648.1995.22020226.x
24. Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educ Res Rev.* 2007;2(2):130-144. DOI: 10.1016/j.edurev.2007.05.002
25. Kant I. *Kritik der reinen Vernunft*. 2. Auflage. Berlin: de Gruyter; 1968.
26. Kirkpatrick JD, Kirkpatrick WK. *Kirkpatrick's Four Levels of Training Evaluation*. Alexandria, VA: ATD Press; 2016.
27. Kiss A, Steiner C, Grossman P, Langewitz W, Tschudi P, Kiessling C. Students' satisfaction with general practitioners' feedback to their reflective writing: a randomized controlled trial. *Can Med Educ J.* 2017;8(4):e54-e59. DOI: 10.36834/cmej.36929
28. Koh YH, Wong ML, Lee JJ. Medical students' reflective writing about a task-based learning experience on public health communication. *Med Teach.* 2014;36(2):121-129. DOI: 10.3109/0142159X.2013.849329
29. Koole S, Dornan T, Aper L, Scherpbier A, Valcke M, Cohen-Schotanus J, Derese A. Factors confounding the assessment of reflection: a critical review. *BMC Med Educ.* 2011;11:104. DOI: 10.1186/1472-6920-11-104
30. Learman LA, Autry AM, O'Sullivan P. Reliability and validity of reflection exercises for obstetrics and gynecology residents. *Am J Obstet Gynecol.* 2008;198(4):461.e1-8. DOI: 10.1016/j.ajog.2007.12.021
31. Lim JY, Ong SY, Ng CY, Chan KL, Wu SY, So WZ, Tey GJ, Lam YX, Gao NL, Lim YX, Tay RY, Leong IT, Rahman ND, Chiam M, Lim C, Phua GL, Murugam V, Ong EK, Krishna LK. A systematic scoping review of reflective writing in medical education. *BMC Med Educ.* 2023;23(1):12. DOI: 10.1186/s12909-022-03924-4
32. Lovell BL, Lee RT. Burnout and health promotion in veterinary medicine. *Can Vet J.* 2013;54(8):790-791.
33. Lutz G, Scheffler C, Edelhaeuser F, Tauschel D, Neumann M. A reflective practice intervention for professional development, reduced stress and improved patient care – a qualitative developmental evaluation. *Patient Educ Couns.* 2013;92(3):337-345. DOI: 10.1016/j.pec.2013.03.020
34. Lyon P, Letschka P, Ainsworth T, Haq I. An exploratory study of the potential learning benefits for medical students in collaborative drawing: creativity, reflection and 'critical looking'. *BMC Med Educ.* 2013;13:86. DOI: 10.1186/1472-6920-13-86
35. Makoul G, Zick AB, Aakhus M, Neely KJ, Roemer PE. Using an online forum to encourage reflection about difficult conversations in medicine. *Patient Educ Couns.* 2010;79(1):83-86. DOI: 10.1016/j.pec.2009.07.027
36. *Medizinischer Fakultätentag (MFT). Nationaler Kompetenzbasierter Lernzielkatalog Medizin*. Berlin: MFT; 2015.
37. Messick S. Validity of test interpretation and use. In: Alkin MC, editor. *Encyclopedia of Educational Research*. 6th ed. New York: Macmillan; 1991.
38. Messick S. Validity. In: Linn RL, editor. *The American Council on Education/Macmillan series on higher education Educational measurement*. New York: Macmillan; 1989. p.13-103.
39. Miller-Kuhlman R, O'Sullivan PS, Aronson L. Essential steps in developing best practices to assess reflective skill: A comparison of two rubrics. *Med Teach.* 2016;38(1):75-81. DOI: 10.3109/0142159X.2015.1034662
40. Moniz T, Arntfield S, Miller K, Lingard L, Watling C, Regehr G. Considerations in the use of reflective writing for student assessment: issues of reliability and validity. *Med Educ.* 2015;49(9):901-908. DOI: 10.1111/medu.12771
41. Moon JA. *A handbook of reflective and experiential learning: Theory and practice*. London: Psychology Press; 2004.
42. Moskal BM. *Developing Classroom Performance Assessments and Scoring Rubrics-Part II*. ERIC Digest. 2003. Zugänglich unter/available from: <https://eric.ed.gov/?id=ED481715>
43. Moskal BM. Scoring rubrics: what, when and how? *Pract Assess Res Eval.* 2000;7(3). Zugänglich unter/available from: <http://pareonline.net/getvn.asp?v=7&n=34>
44. O'Sullivan P, Aronson L, Chittenden E, Niehaus B, Learman L. Reflective ability rubric and user guide. *MedEdPortal.* 2010. DOI: 10.15766/mep\_2374-8265.8133
45. OIE. OIE recommendations on the Competencies of graduating veterinarians ('Day 1 graduates') to assure National Veterinary Services of quality. Paris: OIE; 2012.

46. Ottenberg AL, Pasalic D, Bui GT, Pawlina W. An analysis of reflective writing early in the medical curriculum: The relationship between reflective capacity and academic achievement. *Med Teach.* 2016;38(7):724-729. DOI: 10.3109/0142159X.2015.1112890
47. Pee B, Woodman T, Fry H, Davenport ES. Appraising and assessing reflection in students' writing on a structured worksheet. *Med Educ.* 2002;36(6):575-585. DOI: 10.1046/j.1365-2923.2002.01227.x
48. Pennebaker JW. *Opening up: The healing power of expressing emotions.* London, New York: Guilford Press; 1997.
49. Randolph JJ. Free-marginal multirater kappa (multirater kfree): an alternative to Fleiss' fixed-Marginal multirater kappa. In: Paper presented at the Joensuu Learning and Instruction Symposium, Joensuu, Finland. 2005. Zugänglich unter/available from: [https://www.researchgate.net/publication/224890485\\_Free-Marginal\\_Multirater\\_Kappa\\_multirater\\_kfree\\_An\\_Alternative\\_to\\_Fleiss\\_Fixed-Marginal\\_Multirater\\_Kappa](https://www.researchgate.net/publication/224890485_Free-Marginal_Multirater_Kappa_multirater_kfree_An_Alternative_to_Fleiss_Fixed-Marginal_Multirater_Kappa)
50. Reis SP, Wald HS, Monroe AD, Borkan JM. Begin the BEGAN (The Brown Educational Guide to the Analysis of Narrative) – A framework for enhancing educational impact of faculty feedback to students' reflective writing. *Patient Educ Couns.* 2010;80(2):253-259. DOI: 10.1016/j.pec.2009.11.014
51. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Med Teach.* 2009;31(8):685-695. DOI: 10.1080/01421590903050374
52. Sandars J, Murray C. Digital storytelling for reflection in undergraduate medical education: a pilot study. *Educ Prim Care.* 2009;20(6):441-444. DOI: 10.1080/14739879.2009.11493832
53. Shapiro J, Rakhra P, Wong A. The stories they tell: How third year medical students portray patients, family members, physicians, and themselves in difficult encounters. *Med Teach.* 2016;38(10):1033-1040. DOI: 10.3109/0142159X.2016.1147535
54. Stevens DD, Cooper JE. *Journal Keeping: How to Use Reflective Writing for Learning, Teaching, Professional Insight and Positive Change.* Sterling (VA): Stylus Publishing; 2009.
55. Uygur J, Stuart E, De Paor M, Wallace E, Duffy S, O'Shea M, Smith S, Pawlikowska T. A Best Evidence in Medical Education systematic review to determine the most effective teaching methods that develop reflection in medical students: BEME Guide No. 51. *Med Teach.* 2019;41(1):3-16. DOI: 10.1080/0142159X.2018.1505037
56. Wald HS. Professional identity (trans)formation in medical education: reflection, relationship, resilience. *Acad Med.* 2015;90(6):701-706. DOI: 10.1097/ACM.0000000000000731
57. Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing. *Acad Med.* 2012;87(1):41-50. DOI: 10.1097/ACM.0b013e31823b55fa
58. Wald HS, Haramati A, Bachner YG, Urkin J. Promoting resiliency for interprofessional faculty and senior medical students: Outcomes of a workshop using mind-body medicine and interactive reflective writing. *Med Teach.* 2016;38(5):525-528. DOI: 10.3109/0142159X.2016.1150980
59. Wald HS, Reis SP, Borkan JM. Reflection rubric development: evaluating medical students' reflective writing. *Med Educ.* 2009;43(11):1110-1111.
60. West CP. Physician Well-Being: Expanding the Triple Aim. *J Gen Intern Med.* 2016;31(5):458-459. DOI: 10.1007/s11606-016-3641-2
61. Wong FK, Kember D, Chung LY, Yan L. Assessing the level of student reflection from reflective journals. *J Adv Nurs.* 1995;22(1):48-57. DOI: 10.1046/j.1365-2648.1995.22010048.x

**Corresponding author:**

Dr. Sabine Ramspott  
Trillium GmbH Medizinischer Fachverlag, Jesenwanger  
Str. 42b, D-82284 Grafrath, Germany, Phone: +49  
(0)176/76788212  
sabine.ramspott@trillium.de

**Please cite as**

Ramspott S, Sonntag U, Härtl A, Rüttermann S, Roller D, Giesler M, Hempel L. *MaReS (Magdeburg Reflective Writing Scoring Rubric for Feedback) – development of a feedback method for reflective writing in health professions education: A pilot study in veterinary medicine.* *GMS J Med Educ.* 2025;42(2):Doc28.  
DOI: 10.3205/zma001752, URN: urn:nbn:de:0183-zma0017525

**This article is freely available from**

<https://doi.org/10.3205/zma001752>

**Received:** 2021-10-12

**Revised:** 2024-07-26

**Accepted:** 2024-12-16

**Published:** 2025-04-15

**Copyright**

©2025 Ramspott et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License. See license information at <http://creativecommons.org/licenses/by/4.0/>.

# MaReS (Magdeburger Reflexionsskala) – Entwicklung einer Feedback-Methode für reflektierendes Schreiben in der Ausbildung von Gesundheitsberufen: Eine Pilotstudie in der Veterinärmedizin

## Zusammenfassung

**Zielsetzung:** Das Ziel der Studie war die Entwicklung eines Bewertungsinstrumentes, mit dem Studierenden konstruktives Feedback gegeben werden kann sowie Belege für die Konstruktvalidität des Instruments zu sammeln.

**Methodik:** Die Magdeburger Reflexionsskala (MaReS) wurde in einem iterativen Prozess im Anschluss an einen Workshop zum Thema Reflexion durch eine Arbeitsgruppe eines Ausschusses der „Gesellschaft für Medizinische Ausbildung (GMA)“ im Juni 2016 entwickelt.

25 Berichte von 13 Studierenden der Veterinärmedizin wurden in zwei Durchgängen (13 und zwölf Berichte) von drei unabhängigen Beurteilenden mit MaReS und von zwei Beurteilenden mit dem Instrument REFLECT bewertet. Die Validität wurde anhand der folgenden Komponenten der Konstruktvalidität von Messick ermittelt: Inhalt (Entwickeln des Instruments), Antwortprozess (Kodierleitfaden, Schulung der Beurteilenden, Bewertungsdauer, Evaluation durch Studierende), interne Struktur (Inter-Rater-Reliabilität, IRR) und Beziehung zu anderen Variablen (Vergleich der Bewertung mit Instrument REFLECT und einer globalen Bewertungsskala).

**Ergebnisse:** Das analytische Instrument umfasst zwölf Items, die auf einer dreistufigen Skala bewertet werden. Die Autor:innen entwickelten eine Aufgabenstellung mit Leitfragen für die Studierenden sowie einen Kodierleitfaden. Die Ergebnisse für das freie marginale Kappa der Items von MaReS reichten von -0,08 bis 0,77 für die erste Gruppe von Reflexionsberichten und von 0,13 bis 0,75 für die zweite Gruppe. Die Korrelationen zwischen MaReS und REFLECT waren positiv (erster Durchlauf:  $r=0,92$  ( $p<0,001$ ); zweiter Durchlauf:  $r=0,29$  ( $p=0,37$ )).

**Schlussfolgerung:** MaReS kann ein nützliches Instrument sein, um das reflektierende Schreiben von Studierenden in der Ausbildung von Gesundheitsberufen sowohl anzuleiten als auch strukturiertes Feedback zu geben. Die Verwendung von weiteren Berichten zur Schulung der Beurteilenden und mehr Trainingszyklen würden vermutlich zu höheren IRR führen.

**Schlüsselwörter:** Reflexion, reflektierendes Schreiben, analytisches Bewertungsinstrument, Validität, Feedback

Sabine Ramspott<sup>1</sup>

Ulrike Sonntag<sup>2</sup>

Anja Härtl<sup>3</sup>

Stefan Rüttermann<sup>4</sup>

Doris Roller<sup>5</sup>

Marianne Giesler<sup>6</sup>

Linn Hempel<sup>7</sup>

1 Trillium GmbH Medizinischer Fachverlag, Grafrath, Deutschland

2 Charité – Universitätsmedizin Berlin, Institut für Allgemeinmedizin, Berlin, Deutschland

3 Universitätsklinikum Augsburg, Stabsstelle für Hygiene und Umweltmedizin, Augsburg, Deutschland

4 Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz, Deutschland

5 Ruprecht-Karls-Universität Heidelberg, Zentrum für Psychosoziale Medizin, Institut für Medizinische Psychologie, Heidelberg, Deutschland

6 Freiburg/Br., Deutschland

7 Martin-Luther-Universität Halle-Wittenberg, Medizinische Fakultät, Dorothea Erxleben Lernzentrum, Halle/Saale, Deutschland

# 1. Einleitung und Zielsetzung

## 1.1. Reflexion in der Ausbildung von Gesundheitsberufen

Um die Patientenversorgung zu verbessern, wurden verschiedene Rollen, die Mediziner\*innen [16] erfüllen müssen, sowie eine Reihe von Domänen für die tierärztliche Praxis [5] definiert. Um den Anforderungen dieser Rollen bzw. Domänen gerecht zu werden, müssen Fachkräfte im Gesundheitswesen während ihres Studiums zahlreiche entsprechende Kompetenzen erwerben [https://www.avma.org/education/center-for-veterinary-accreditation/accreditation-policies-and-procedures-avma-council-education-coe], [13], [16], [36], [45]. Da die Anforderungen und das Wissen stetig wachsen, wird das Arbeitsumfeld der Fachkräfte immer komplexer. Um eine zeitgemäße Gesundheitsversorgung zu gewährleisten, müssen Mediziner\*innen und Tiermediziner\*innen lebenslang lernen, wobei die Fähigkeit zur Reflexion des eigenen Handelns ein wesentlicher Bestandteil ist [29]. Reflexion wurde bereits in der Aufklärung (vgl. 18. Jh. Kant, Fries u. a.) als wesentliches Mittel der Erkenntnis angesehen und u. a. als „innere Einkehr“ beschrieben [17], [25]. In der heutigen Literatur wird Reflexion u. a. als „in-depth consideration of events or situations: the people involved, what they experienced, and how they felt about it“ [6] beschrieben.

Reflexion unterstützt somit auch den Prozess der beruflichen Identitätsbildung, einen wichtigen Entwicklungsprozess für Personen in den Gesundheitsberufen [9]. Darüber hinaus zeigt sich, dass Reflexion die Resilienz zu fördern [56] und Stress zu reduzieren scheint [33]. Diese Erkenntnisse werden auch deswegen relevanter, da das Wohlbefinden von Mediziner\*innen und Tiermediziner\*innen in den letzten Jahren immer mehr an Bedeutung gewonnen hat [19], [31], [32], [55], [60].

Die Fähigkeit zur Reflexion kann als Teil eines Lernprozesses [54] entwickelt und durch verschiedene Aktivitäten gefördert werden: Reflektierendes Schreiben ist in der medizinischen Ausbildung weit verbreitet [27], [28], [46], [53], [58].

Andere Aktivitäten, die das Reflektieren fördern, sind Gruppendiskussionen [7], [33], Online-Foren [35], „digital Storytelling“ [52] und gemeinsames Zeichnen [34].

„Schreiben als Methode“ ist nachweisbar wirksam [48]. Indem Menschen eine wahrgenommene Situation aufschreiben, werden sie sich des Geschehenen gewahr und trainieren zusätzlich ihre Ausdrucksfähigkeit. Dies wirkt sich auf das Denken, Fühlen und Handeln aus. Am meisten profitieren Menschen, die im Alltag wenig Gelegenheit zur Selbstreflexion haben, denen das Erkennen und Benennen von Gefühlen schwerfällt und die im Schreiben Perspektiven zu ihren festen Denkmustern ausprobieren können [54].

Um Reflexion zu fördern, kann Feedback ein äußerst relevantes und wirkungsvolles Instrument sein. Ein systematisches Review zeigte jedoch, dass in den meisten

Studien, die Feedback als Element ihrer Intervention zur Förderung der Reflexion einsetzten, die Methoden und Feedbackdetails unklar blieben [55]. Nur eine der Studien beschrieb das Protokoll, das für die Bereitstellung von Feedback verwendet wurde [3].

Im deutschsprachigen Raum gibt es viele große medizinische und veterinärmedizinische Fakultäten mit bis zu 870 Studienanfänger\*innen (z. B. Ludwig-Maximilians-Universität München, Deutschland). Reflektierendes Schreiben scheint ein ideales Mittel zu sein, um eine große Anzahl von Studierenden in Reflexionsprozesse zu involvieren. Es ist jedoch ressourcenintensiv, Reflexion zu lehren und einer großen Anzahl von Studierenden Feedback zu ihren Texten zu geben. Ein individuelles schriftliches Feedback ist sehr wahrscheinlich zeitaufwändiger als die Nutzung eines Instruments, mit dem durch Ankreuzen Feedback gegeben werden kann. Somit können entsprechende Bewertungsinstrumente ein geeignetes Mittel sein, um das Geben von Feedback zu vereinfachen. Da der Einsatz von strukturierter Reflexion in der Ausbildung von Gesundheitsberufen in Deutschland, Österreich und der Schweiz (DACH-Region) relativ neu ist, sind Lehrende in der Regel weder in der Vermittlung dieser Kompetenz noch im Geben von Feedback besonders erfahren.

## 1.2. Bewertungsinstrumente

Zur Bewertung von Leistungen und schriftlichen Reflexionen werden häufig spezielle Instrumente verwendet. Es gibt zwei verschiedene Arten der Bewertung: Bei der ganzheitlichen Bewertung bewerten die Prüfenden den Aufsatz des Studierenden anhand einer vorgegebenen Skala global, während bei der analytischen Bewertung der Aufsatz in einzelne Elemente zerlegt wird, von denen jedes einen Anteil der verfügbaren Note erhält [22]. Auszubildende verwenden in der Regel die ganzheitliche Bewertung für umfangreiche Beurteilungen, da sie als einfach zu handhaben, kostengünstig und genau gilt. Die analytische Bewertung ist für die Lehre nützlich, da die Ergebnisse Lehrenden und Studierenden helfen können, die Stärken und Lernbedürfnisse der Studierenden zu erkennen [24]. Soll die Bewertung eines Reflexionsberichts als formative Beurteilung verwendet werden, scheint der analytische Ansatz besser geeignet.

Lehrende in der Medizin haben mehrere holistische [30], [44], [47], [51], [59], [61] und analytische [10], [11], [57] Instrumente für die formative und/oder summative Bewertung von Reflexionsberichten entwickelt. Reis et al. entwickelten „The Brown Educational Guide to the Analysis of Narrative“ (BEGAN), einen Leitfaden für die Erstellung von schriftlichem Feedback zu Reflexionsberichten von Studierenden [50]. Dieser Leitfaden und die analytischen Instrumente können verwendet werden, um Feedback zu den Reflexionsberichten der Studierenden zu geben. Das Instrument REFLECT umfasst fünf Hauptkriterien (Schreibspektrum, Präsenz, Beschreibung eines Konflikts oder desorientierenden Dilemmas, Beachtung von Emotionen, Analyse und Beimessung von Bedeutung)

sowie ein optionales Nebenkriterium (Beachtung der Aufgabenstellung). Diesen Kriterien sind vier Reflexionsstufen zugeordnet (gewöhnheitsmäßiges Handeln, überlegtes Handeln oder Introspektion, Reflexion und kritische Reflexion). Wenn die Stufe der kritischen Reflexion erreicht wird, sind die Lernergebnisse auch als transformatives oder bestätigendes Lernen zu definieren [57]. Das Instrument von Devlin et al. umfasst vier Dimensionen (deskriptiv, vergleichend, persönlich und kritisch) mit Fragen, die die Bewertung und das Feedback anleiten [11]. Das Instrument von Devi basiert auf den Indikatoren von Koole [29] zur Beschreibung des Reflexionsprozesses und dem Benotungssystem von Moon [41]. Die Bewertenden benoten die Berichte von A bis F [10]. Bislang konnte kein Bewertungsinstrument für reflektierendes Schreiben im medizinischen Kontext in deutscher Sprache identifiziert werden.

### 1.3. Leitfäden für reflektierendes Schreiben

Es gibt mehrere Leitfäden für das reflektierende Schreiben in der medizinischen Ausbildung. Im Anhang von Sandars [51] AMEE-Leitfadens finden sich verschiedene Materialien: Beispiele für ein Informationsblatt für Medizinstudierende im Grund- und Aufbaustudium, das auf Moons Handbuch [41] basiert, eine Vorlage für eine strukturierte Reflexion zur Entwicklung einer therapeutischen Beziehung in der beruflichen Praxis nach Johns [23] mit 20 Leitfragen sowie elf Fragen, um die Reflexion zu vertiefen. Aronson et al. [4] entwickelten einen Leitfaden für reflexives Lernen, der aus einem strukturierten Ansatz zur Reflexion auf der Grundlage des SOAP-Formats mit Leitfragen und einem Informationsblatt über Reflexion und Strategien für erfolgreiche Reflexion besteht. Aronson et al. testeten ihr Instrument anhand eines holistischen Bewertungstools [4]. Die Autor\*innen dieses Artikels konnten weder ein analytisches Bewertungsinstrument mit dazugehörigen Leitfragen noch einen Leitfaden für Studierende finden.

### 1.4. Zielsetzung

Ziel der Studie war es, ein deutsches Bewertungsinstrument für die formative Beurteilung von Reflexionsberichten zu entwickeln, das sich an Leitfragen für Studierende orientiert und ihnen wertvolle Rückmeldungen gibt. Die Nutzung des Instruments soll für Lehrende bei der Bewertung von Reflexionsberichten einfach und zeitsparend sein. Darüber hinaus war es das Ziel der Autor\*innen, Belege für die Validität des neuen Instruments zu sammeln.

## 2. Methoden

### 2.1. Entwicklung des Bewertungsinstruments

Der Ausschuss "Kommunikative und soziale Kompetenzen" (KusK) der Gesellschaft für Medizinische Ausbildung (GMA) veranstaltete im Juni 2016 ein Workshop zum Thema Reflexion, um Möglichkeiten zu erörtern, Reflexion in die Lehre von kommunikativen und sozialen Kompetenzen für Medizin- und Tiermedizinstudierende in Deutschland zu integrieren. Die Teilnehmenden eines Workshops untersuchten drei bestehende Instrumente zur Bewertung von Reflexionsberichten [44], [50], [57]. Basierend auf diesen drei Instrumenten wurde eine erste Version eines Bewertungsinstruments entwickelt.

Nach dem Workshop wurde das Bewertungsinstrument präzisiert, wobei Kooles Modell gemeinsamer Elemente zur Beschreibung eines Reflexionsprozesses [29] als Leitfaden für die Auswahl der Inhalte diente. Hinsichtlich der Struktur und des Wortlauts des Bewertungsinstruments wurde den Empfehlungen von Moskal [43] zur Entwicklung von Bewertungsinstrumenten gefolgt. Das Feedbackinstrument wurde zusammen mit einer passenden Aufgabenstellung, Leitfragen für Studierende sowie einem Kodierleitfaden in einem iterativen Prozess entwickelt. Nach mehreren Überarbeitungen und einem ersten Konsensverfahren durch die Autor\*innen dieser Publikation wurden das Instrument, die Aufgabenstellung und die Leitfragen, der Kodierleitfaden und ein Beispiel-Reflexionsbericht an sechs externe Expert\*innen geschickt, die sich an medizinischen und veterinärmedizinischen Fakultäten in Deutschland mit Reflexion in der Lehre befassen. Die Expertengruppe setzte sich aus Mediziner\*innen, Tiermediziner\*innen, Gesundheitswissenschaftler\*innen und Psycholog\*innen zusammen. Deren Rückmeldungen zum Instrument und seinen Items wurden diskutiert und nach einem Konsens in die neue Version des Instruments eingearbeitet. Das Instrument wurde in Anlehnung an seinen Entstehungsort „Magdeburger Reflexionsskala“ (MaReS) genannt.

### 2.2. Belege für die Konstruktvalidität

„Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment“ [37]. Das Bewertungsinstrument selbst kann nicht für valide (oder nicht valide) erklärt werden, aber es können mehr (oder weniger) Belege für seine Validität gesammelt werden, um die vorgeschlagenen Interpretationen der Bewertungsergebnisse zu unterstützen. Der Kontext, in dem die Belege für die Validität gesammelt werden, ist nach wie vor wichtig: Bewertungsergebnisse sind für einen bestimmten Zweck, eine bestimmte Bedeutung oder Interpretation zu einem bestimmten Zeitpunkt und für eine bestimmte Population, für die

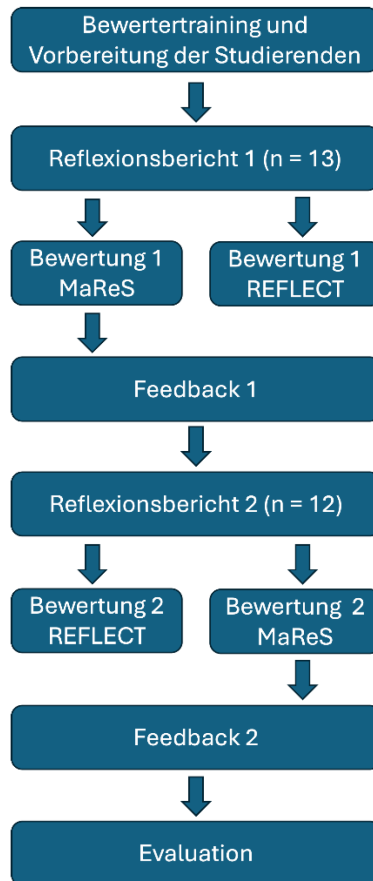


Abbildung 1: Studiendesign MaReS

<b>Inhalt</b>	<ul style="list-style-type: none"> <li>• Nutzung von Kooles Modell von gemeinsamen Elementen der Reflexion</li> <li>• Nutzung von Moskals Empfehlungen für das Entwickeln von Bewertungsinstrumenten</li> <li>• Diskussion mit Expert*innen</li> </ul>
<b>Antwortprozess</b>	<ul style="list-style-type: none"> <li>• Entwicklung und Nutzung eines Bewerterhandbuchs und Bewertertrainings</li> <li>• Beurteilungsdauer</li> <li>• Evaluation durch die Studierenden</li> </ul>
<b>Interne Struktur</b>	<ul style="list-style-type: none"> <li>• Inter-Rater-Reliabilität</li> </ul>
<b>Bezug zu anderen Variablen</b>	<ul style="list-style-type: none"> <li>• Vergleich mit einem anderen Instrument (REFLECT und eine global Bewertungsskala)</li> </ul>

Abbildung 2: Quellen für Belege der Validität von MaReS

Belege für die Validität gesammelt wurden, mehr – oder weniger – gültig [12]. In dieser Studie haben wir Belege für die Validität für die Verwendung des Instruments als formatives Beurteilungsinstrument der Reflexionsfähigkeit in Reflexionsberichten gesammelt. Die Berichte über die tierärztliche Praxis wurden von Tiermedizinstudierenden verfasst, die 2017 an einem Wahlpflichtfach zum Thema Reflexion teilnahmen und von fünf Bewertenden bewertet, die die Studierenden nicht kannten. Jeder Bericht wurde von drei Bewertenden mit dem neu entwickelten Bewertungsinstrument und von zwei mit dem Instrument REFLECT bewertet. Die Autor\*innen des vorliegenden Artikels sammelten Belege für die Validität für vier der sechs

von Messick [38] aufgeführten Quellen der Konstruktvalidität: Inhalt, Antwortprozess, interne Struktur und Bezug zu anderen Variablen. Abbildung 1 zeigt das Studiendesign; Abbildung 2 zeigt die für diese Studie ausgewählten Quellen der Belege für die Validität. Im folgenden Abschnitt beschreiben wir die Methoden, die wir zum Sammeln von Belegen für die verschiedenen Quellen der Konstruktvalidität verwendet haben in chronologischer Reihenfolge. Daher wird die Quelle „Antwortprozess“ in diesem Abschnitt aufgeteilt.



### 2.2.1. Inhalt – Entwicklung des Bewertungsinstruments

Die Art und Weise, wie wir das Instrument entwickelt haben, trägt zum Nachweis der Inhaltsvalidität bei (siehe 2.1.).

### 2.2.2. Antwortprozess I – Nutzung des Kodierleitfadens und Schulung der Bewertenden

Alle fünf Bewertende nahmen am zweitägigen Kusk-Workshop zu Reflexion im Juni 2016 teil. Nach der Entwicklung von MaReS wurde den Bewertenden (us, lh, srü, dr, ah) die endgültige Version des Instruments, die Aufgabenstellung, die Leitfragen und der Kodierleitfaden zur Verfügung gestellt. Sie bewerteten jeweils einzeln und unabhängig den gleichen Beispielbericht. Die Bewertungen wurden verglichen und abweichende Bewertungen wurden mit allen Bewertenden in einer Telefonkonferenz diskutiert. Anschließend einigten sich alle Bewertenden auf eine gemeinsame Vorgehensweise bezüglich der Items. Der Kodierleitfaden wurde entsprechend angepasst.

Der gleiche Ansatz wurde für das Instrument REFLECT [58] verfolgt, das zur Bewertung des Bezugs zu anderen Variablen verwendet wurde (siehe 2.3.). Da hier zu diesem Zeitpunkt kein Kodierleitfaden vorlag, war eine Anpassung nicht erforderlich. Somit wurden die englische Version von REFLECT und die vier Schritte für seine Anwendung verwendet [57].

### 2.2.3. Interne Struktur – Inter-Rater-Reliabilität und Bezug zu anderen Variablen – Vergleich mit einem anderen Bewertungsinstrument

Die Ethikkommission der LMU München erteilte das Ethikvotum für die Studie (Projektnummer 17-065). Dreizehn Studierende eines Wahlpflichtfachs über Reflexion in der tierärztlichen Praxis schrieben jeweils einen Reflexionsbericht, nachdem sie sechs Stunden Unterricht zu Reflexion im Allgemeinen sowie zu verschiedenen Aktivitäten zur Förderung von Reflexion, einschließlich des reflektierenden Schreibens auf Grundlage von MaReS erhalten hatten. Die Studierenden wurden mit dem Bewertungsinstrument vertraut gemacht. Sie diskutierten verschiedene Beispiele für reflektierendes Schreiben und die Bewertung mit MaReS. Die Aufgabe bestand darin, einen Reflexionsbericht über eine konkrete Situation zu schreiben. Die Studierenden sollten eine Situation aus ihrem Studium oder Beruf wählen, in der sie sich in der Interaktion mit Patient\*innen, Patientenbesitzer\*innen, Kommiliton\*innen oder Lehrenden herausgefordert fühlten. Die 13 Berichte wurden von drei Bewertenden mit MaReS (ah, dr, lh) und von zwei Bewertenden mit REFLECT (us, srü) bewertet. Eine der fünf MaReS-Bewertungen wurde an die Studierenden zurückgemeldet. Anschließend schrieben zwölf der Studierenden einen wei-

teren Reflexionsbericht, der wiederum von drei Bewertenden (us, srü, lh) mit MaReS und von zwei Bewertenden (ah, dr) mit REFLECT bewertet wurde. Wieder wurde eine der MaReS-Bewertungen an die Studierenden zurückgemeldet. Alle Bewertenden füllten für jeden Bericht eine fünfstufige Gesamtwertung aus. Die Gesamt- und REFLECT-Bewertung wurde nur für Studienzwecke verwendet und nicht an die Studierenden zurückgemeldet.

### 2.2.4. Antwortprozess II – Bewertungsdauer und Notizen der Bewertenden

Die Bewertenden protokollierten auch die Zeit, die sie für ihre Bewertung benötigten und machten, falls erforderlich, Notizen zu wichtigen Aspekten des Berichts, protokollierten schwierige Items und mögliche neue Ankerbeispiele für den Kodierleitfaden.

### 2.2.5. Antwortprozess III – Evaluation durch die Studierenden

Nachdem sie zwei MaReS-Feedbackbögen zu ihren Berichten erhalten hatten, füllten acht Studierende einen Fragebogen aus. Er umfasste fünf allgemeine Fragen zur Reflexion und zum Wahlpflichtfach, sechs Fragen zur Aufgabe und zu den Leitfragen, vier Fragen zu den erhaltenen Rückmeldungen und fünf Fragen zu ihrer Reflexionsfähigkeit und ihren Ansichten über reflektierendes Schreiben.

## 2.3. Statistische Methoden

Die Bewertungszeit für die Berichte wurde mittels deskriptiver Statistik unter Verwendung von Mittelwerten und Standardabweichung (SD) ausgewertet. Die Inter-Rater-Reliabilität (IRR) wurde mit dem freien marginalen Kappa [49] ermittelt. Die Korrelation zwischen der MaReS-Gesamtnote (Summe der MaReS-Gesamtnote von drei Bewertenden), der REFLECT-Gesamtnote (Summe der REFLECT-Gesamtnote von zwei Bewertenden) und der Gesamtnote (Summe der entsprechenden Gesamtnote von drei Bewertenden für MaReS und von zwei Bewertenden für REFLECT) wurde mit Hilfe des Spearmans Rangkorrelationskoeffizienten ermittelt.

Die Evaluation durch die Studierenden und das Feedback der Bewertenden wurden mittels deskriptiver Statistik (absolute Zahlen) für Fragen mit Antworten auf einer Fünfpunkte-Skala und mittels Inhaltsanalyse nach Mayring für offene Fragen analysiert.

## 3. Ergebnisse

### 3.1. Entwicklung des Bewertungsinstruments

Wir entwickelten eine Schreibaufgabe mit neun Leitfragen sowie ein Instrument mit zwölf Items, die mit einer Drei-

**Tabelle 1: Inter-Rater-Reliabilität für MaReS und REFLECT für die verschiedenen Bewertenden**

Item/ Kriterium	freies marginales Kappa			
	MaReS		REFLECT	
	AH, DR, LH	US, SRÜ, LH	US, SRÜ	AH, DR
1	<b><u>0,77</u></b>	<b><u>0,75</u></b>	0,28	0,22
2	<b><u>0,77</u></b>	<b><u>0,67</u></b>	0,08	-0,11
3	<b><u>0,62</u></b>	<b><u>0,75</u></b>	0,18	0,33
4	<b><u>0,65</u></b>	0,58	-0,26	0,00
5	0,39	0,25	0,28	0,22
6	0,08	<b><u>0,75</u></b>	-0,03	0,22
7	-0,08	0,13		
8	0,19	0,13		
9	<b><u>0,62</u></b>	0,42		
10	0,04	0,25		
11	0,35	0,33		
12	0,08	0,25		
		transformatives vs. bestätigendes Lernen	0,31	0,38
		Gesamtbewertung	0,18	0,22

Werte zwischen 0,60 und 0,74 gelten als „gut“ und sind durch fettgedruckte Zahlen gekennzeichnet.

Werte zwischen 0,75 und 1 gelten als „ausgezeichnet“ und sind durch fettgedruckte und unterstrichene Zahlen gekennzeichnet.

Punkte-Skala bewertet wurden (siehe Anhang 1). Der Kodierleitfaden besteht aus einer kurzen Beschreibung, wie der Bericht zu bewerten ist, sowie aus Deskriptoren und Ankerbeispielen für alle Items und konsekutiven Skalen.

## 3.2. Belege für die Konstruktvalidität

### 3.2.1. Inhalt – Entwicklung von MaReS

Siehe 3.1.

### 3.2.2. Antwortprozess – Bewertungsdauer

Die Zeichenzahl (einschließlich Leerzeichen) der Berichte reichte von 2.765 bis 8.488 mit einem Mittelwert von 5.359 Zeichen (SD 1.660). Die durchschnittliche Zeit, die für die Bewertung eines Berichts mit MaReS benötigt wurde, betrug 13,9 Minuten mit einer Standardabweichung von 10,9. Die mittlere Bewertungszeit, d. h. die Zeit, die zum Lesen und Bewerten des Berichts benötigt wurde, variierte stark zwischen den Bewertenden und reichte von durchschnittlich 6,7 Minuten (SD 2,5) bis 28,5 Minuten (SD 13,1). Die mittlere Dauer der REFLECT-Bewertungen betrug 7,8 Minuten mit einer Standardabweichung von 5,2. Die mittleren Bewertungszeiten für die einzelnen Bewertenden reichten von 4,5 Minuten (SD 1,2) bis 14,5 (SD 6,9).

### 3.2.3. Interne Struktur – Inter-Rater-Reliabilität

Die Ergebnisse für das freie marginale Kappa der MaReS-Items reichten von -0,08 bis 0,77 für die erste Runde von Reflexionsberichten und von 0,13 bis 0,75 für die zweite Runde (siehe Tabelle 1). Die Ergebnisse für das freie marginale Kappa der einzelnen Kriterien von REFLECT reichten von -0,26 bis 0,31 für die erste Runde von Reflexionsberichten und von -0,11 bis 0,38 für die zweite Runde (siehe Tabelle 2). Das freie marginale Kappa für die Gesamtbewertung aller fünf Bewertenden betrug 0,16 für die erste Runde von Berichten und 0,22 für die zweite Runde.

### 3.2.4. Bezug zu anderen Variablen – Korrelation der Bewertungen

Für die Berichte 1 bis 13 betrug Spearmans Rangkorrelationskoeffizient für MaReS und die entsprechende Gesamtbewertung der Bewertenden  $r=0,43$  ( $p=0,14$ ),  $r=0,92$  ( $p<0,001$ ) für MaReS und REFLECT und  $r=0,83$  ( $p<0,001$ ) für REFLECT und die Gesamtbewertung.

Für die Berichte 14 bis 25 betrug Spearmans Rangkorrelationskoeffizient für MaReS und die entsprechende Gesamtbewertung der Bewertenden  $r=0,75$  ( $p=0,005$ ),  $r=0,29$  ( $p=0,37$ ) für MaReS und REFLECT und  $r=0,87$  ( $p<0,001$ ) für REFLECT und die entsprechende Gesamtbewertung der Bewertenden.

**Tabelle 2: Antworten auf die geschlossenen Fragen in den Evaluationen der Studierenden**

Frage		Ausprägung (Anzahl der Antworten)				
Wie würden Sie Ihre „Schreibkompetenz“, unabhängig von den Reflexionsberichten, einschätzen?		sehr gut (1)	eher gut (4)	ganz ok (3)	nicht so gut (0)	nicht gut (0)
Die Aufgabenstellung für die schriftliche Reflexion war für mich konkret und verständlich.		voll und ganz (6)	größtenteils (2)	teils-teils (0)	nicht so richtig (0)	gar nicht (0)
Wie würden Sie die „Schwierigkeit“ der Aufgabe – also das Verfassen des Reflexionsberichts zu einer von Ihnen gewählten herausfordernden Situation – für sich einschätzen?		sehr einfach (1)	eher einfach (2)	teils-teils (4)	eher schwierig (1)	sehr schwierig (0)
Wie beurteilen Sie jetzt im Nachhinein Ihre Auswahl der Situationen für die Reflexionsberichte?	Aufsatz 1	genau passend (4)	passend (2)	teilweise passend (2)	nicht so passend (0)	nicht passend (0)
	Aufsatz 2	genau passend (0)	passend (1)	teilweise passend (5)	nicht so passend (2)	nicht passend (0)
Wie hilfreich waren die Leitfragen für das Verfassen der Reflexionsberichte für Sie?		sehr hilfreich (7)	hilfreich (1)	teilweise hilfreich (0)	nicht so hilfreich (0)	gar nicht hilfreich (0)
Gab es Momente, während des Verfassens des Berichts, an denen Sie „einfach nicht mehr weiter gekommen“ sind, das Schreiben sehr schwierig war oder/und Sie gezögert haben, zu schreiben, was Sie denken/fühlen?		ja, immer wieder (0)	ja, manchmal (3)	nur ganz selten (0)	nein, nie (5)	-
Inwiefern hat sich das Wissen, dass Sie Feedback auf Ihren Reflexionsbericht erhalten werden beim Schreiben der Reflexionsberichte eher förderlich ausgewirkt oder behindert?		förderlich (4)	eher förderlich (3)	teils-teils (1)	eher hinderlich (0)	hinderlich (0)
Haben Sie sich beim Verfassen des zweiten Reflexionsberichtes eher sicherer oder nicht sicherer gefühlt als beim Verfassen des ersten Reflexionsberichtes?		deutlich sicherer (0)	etwas sicherer (4)	unverändert (0)	eher unsicherer (4)	unsicherer (0)
Wie würden Sie selbst Ihre eigene Reflexionsfähigkeit einschätzen?		sehr gut (0)	eher gut (4)	ganz ok (3)	nicht so gut (1)	nicht gut (0)

### 3.2.5. Antwortprozess – Evaluation durch die Studierenden

Acht Studierende nahmen an der Evaluation teil. Die Ergebnisse der geschlossenen Fragen sind in Tabelle 2 dargestellt. Alle Studierenden bewerteten ihre eigene Schreibkompetenz positiv und verstanden die Aufgabe. Die Schwierigkeit der Aufgabe („Verfassen eines Reflexionsberichts über eine selbst gewählte herausfordernde Situation“) wurde von den Studierenden unterschiedlich bewertet. Für den ersten Reflexionsbericht bewerteten die meisten Studierenden ihre eigene Situationsauswahl als „perfekt geeignet“ oder „geeignet“. Für den zweiten Reflexionsbericht wurde die Auswahl von den meisten Studierenden als „teilweise geeignet“ bewertet. Alle acht Studierenden bewerteten die Leitfragen als „sehr hilfreich“ oder „hilfreich“. Das Wissen, dass sie ein

Feedback zu ihrem Reflexionsbericht erhalten werden, wurde von den meisten als „förderlich“ oder „eher förderlich“ bewertet. Beim Verfassen des zweiten Reflexionsberichts fühlte sich die Hälfte der Studierenden „etwas sicherer“, die andere Hälfte „etwas unsicherer“. Die Studierenden bewerteten ihre eigene Reflexionsfähigkeit überwiegend als gut (vier Studierende). Vier Studierende gaben an, dass sie die zweite Reflexion als schwieriger empfanden, weil sie Probleme hatten, eine geeignete Situation für die Reflexion zu finden.

Das strukturierte Feedback wurde von fünf Studierenden positiv bewertet. Sie schrieben, dass das Feedback konstruktiv, hilfreich und nicht wertend war. Gleichzeitig kritisierte eine Person, dass sich das Feedback nur auf den Prozess der Reflexion und nicht auf die Situation bezog. In Bezug auf ihre eigene Reflexionsfähigkeit gaben fünf Studierende an, dass sie ihrer Meinung nach sehr

**Tabelle 3: Freitext-Antworten über die Auswirkungen des Schreibens und des Feedbacks auf die Reflexionsberichte**

<b>Inwiefern haben Sie das Schreiben der Reflexionsberichte und das Feedback bei der Verbesserung Ihrer eigenen Selbstwahrnehmung bezüglich der geschilderten Situationen und vielleicht auch darüber hinaus unterstützt?</b>	
Studierende*r 1	Man denkt mehr über das Geschehene und Verbesserungen nach.
Studierende*r 2	Ich habe dadurch die Situationen verarbeitet und gelernt zu reflektieren
Studierende*r 3	Kaum, eigentlich nur Gedanken in Worte gefasst
Studierende*r 4	Längeres und intensiveres Gedankenmachen
Studierende*r 5	-
Studierende*r 6	Es hat mir geholfen die Situation etwas klarer zu sehen. Allerdings sind für mich Gespräche wichtiger als Reflexionsart.
Studierende*r 7	Stark
Studierende*r 8	Ich habe mich durch das Schreiben gezwungen gefühlt, ein Problem anzugehen, das mich sehr belastet, so auch getraut, mich überhaupt selbst wahrzunehmen.
<b>Inwiefern haben Sie das Schreiben der Reflexionsberichte und das Feedback bezüglich Ihrer Sicht auf den Umgang mit Ihren eigenen Emotionen und der Konkretisierung von realistischen Handlungsplänen für ähnliche Situationen und vielleicht auch darüber hinaus unterstützt?</b>	
Studierende*r 1	Ich denke nun mehr über Gespräche mit anderen und über die Gefühle anderer nach.
Studierende*r 2	Es hat mir die Augen aufgemacht; (da ich auch die Perspektive d. anderen betrachten musste). Ich darf nicht so sensibel sein!
Studierende*r 3	-
Studierende*r 4	Sehr!
Studierende*r 5	-
Studierende*r 6	-
Studierende*r 7	Stark -> bin mir bewusster geworden wie ich mit solchen Situationen umgehen kann
Studierende*r 8	siehe oben (Ich habe mich durch das Schreiben gezwungen gefühlt, ein Problem anzugehen, das mich sehr belastet, so auch getraut, mich überhaupt selbst wahrzunehmen.
<b>Was war für Sie insgesamt das relevanteste Lernergebnis, bzw. welche Erkenntnis(se) waren für Sie wirklich gewinnbringend, die Sie durch das Schreiben der Reflexionsberichte und/oder das Feedback erlangen konnten?</b>	
Studierende*r 1	Das ich durch Reflektieren Konflikten aus dem Weg gehen kann
Studierende*r 2	siehe oben (Es hat mir die Augen aufgemacht; (da ich auch die Perspektive d. anderen betrachten musste). Ich darf nicht so sensibel sein!)
Studierende*r 3	Stressbewältigung
Studierende*r 4	Man achtet darauf, sich in einer künftigen Situation anders zu verhalten
Studierende*r 5	-
Studierende*r 6	Prognosen für die Handlungen in der Zukunft zu finden
Studierende*r 7	Reflexion zu machen -> aktiv Fehler anzuerkennen -> nicht Dramatik einsehen
Studierende*r 8	Mir war vorher nie bewusst, wie wichtig es ist, externe Quellen zu beziehen.

selbstkritisch sind und dies die Reflexion erschwert. Andere Antworten waren: nicht in der Lage zu sein, sofort eine Lösung zu finden; mehr Meinungen von anderen zu benötigen; mehr Zeit zum Nachdenken zu haben.

Die Inhaltsanalyse ergab Folgendes: In ihren Freitextantworten nannten mehrere Studierende Stressbewältigung/-abbau als erwarteten persönlichen Nutzen der Reflexion. Darüber hinaus wurden die persönliche Entwicklung in verschiedenen Bereichen, ein besseres Verständnis für und eine bessere Kommunikation mit anderen (auch privat) sowie Vorteile für künftiges Handeln im Allgemeinen und in Lernsituationen genannt.

Darüber hinaus sind alle Freitextantworten auf drei Fragen zu den Auswirkungen des Schreibens von Reflexionsberichten und des Feedbacks dazu in Tabelle 3 zu finden.

## 4. Diskussion

### 4.1. Entwicklung des Bewertungsinstruments

Bei MaReS stand die Praktikabilität im Vordergrund. Ziel war es, ein Bewertungsinstrument zu entwickeln, das einfach zu handhaben ist und den Lehrenden Zeit spart. Die schlechtere Inter-Rater-Reliabilität für REFLECT kann ein Hinweis darauf sein, dass MaReS einfacher zu handhaben ist. Die durchschnittliche Bewertungszeit für MaReS war länger als für REFLECT. Dies spiegelt vermutlich die Tatsache wider, dass MaReS eine größere Anzahl von Items (n=12) aufweist als REFLECT (n=6). Unserer

Meinung nach sind weniger als 15 Minuten (mittlere Bewertungszeit) zum Lesen und Bewerten eines Reflexionsberichts immer noch als angemessen zu einzuordnen.

#### 4.2 Belege für die Konstruktvalidität

In dieser Studie haben wir Belege für die Validität gesammelt, die sich auf die folgenden Komponenten der Konstruktvalidität beziehen: Inhalt, Antwortprozess, interne Struktur und Bezug zu anderen Variablen [1], [12], [38]. MaReS kann ein geeignetes Instrument für das Lehren von Reflexion und das Feedback zu den Reflexionsberichten der Studierenden sein. Die Kombination von Leitfragen für die schriftliche Reflexion und der Einsatz von MaReS als zeitsparende Möglichkeit, strukturiertes Feedback zu geben, wurde erfolgreich pilotiert.

##### 4.2.1. Inhalt

Belege für die Validität, die auf Testinhalten basieren, ergeben sich aus der Analyse der Beziehung zwischen einem Testinhalt und dem Konstrukt, das es messen soll [1]. Der Inhalt von MaReS (einschließlich der Leitfragen, des Bewertungsinstruments und des Kodierleitfadens) wurde sorgfältig auf Grundlage eines allgemein akzeptierten Modells für das Konzept des Reflexionsprozesses ausgewählt [29]. Wir berücksichtigten auch den Inhalt von drei bestehenden Instrumenten zur Bewertung von Reflexionsberichten [44], [50], [57]. Weitere Belege für die Validität wurden durch die Befragung von sechs externen Expert\*innen gewonnen, die den Zusammenhang zwischen Teilen des Tests und dem Konstrukt beurteilen sollten [1]. Ihre Rückmeldungen flossen in die endgültige Version von MaReS ein.

##### 4.2.2. Antwortprozess

Belege für die Validität, die auf dem Antwortprozess basieren, können sowohl den Antwortprozess der Bewertenden als auch den Antwortprozess der Studierenden, die die Aufgabe bearbeiten, umfassen [1].

Um den Bewertungsprozess zu unterstützen, haben wir einen Kodierleitfaden entwickelt, der eine kurze Beschreibung der Bewertung des Berichts sowie Deskriptoren und Ankerbeispiele für alle Items und konsekutiven Skalen enthält. Die Schulung der Bewertenden umfasste einen zweitägigen Workshop zur Reflexion, die Bewertung eines exemplarischen Berichts und eine Diskussion über abweichende Ergebnisse der Bewertung. Am Ende der Diskussion wurde eine Einigung darüber erzielt, wie die Berichte zu bewerten sind.

Obwohl die Inter-Rater-Reliabilität durch die Schulung der Bewertenden verbessert wird, indem man ihnen einen Leitfaden für die Aufzeichnung ihrer Beobachtungen an die Hand gibt [14] und wir in dieser Studie darauf geachtet haben, dass alle Bewertenden verstanden haben, wie die Aufsätze zu bewerten sind, haben wir bei den meisten Items eine schlechte Inter-Rater-Reliabilität festgestellt (siehe 4.2.3.). Dies könnte darauf hindeuten, dass die Schulung der Bewertenden möglicherweise nicht ausreichend war.

Während des Beantwortungsprozesses wurden Daten über die Bewertungsdauer gesammelt. Dies trägt nicht zur Validitätsevidenz bei, sondern soll die Lehrenden in den Gesundheitsberufen, die das reflektierende Schreiben in ihre Curricula aufnehmen möchten, dabei unterstützen, die Durchführbarkeit des Vorhabens zu beurteilen und Ressourcen zu planen.

Zu den Belegen der Validität, die auf dem Antwortprozess der Studierenden basieren, gehört die Vertrautheit der Studierenden mit dem Format [12]. Bevor sie ihre Berichte schrieben, wurden den Studierenden die Leitfragen und das Bewertungsinstrument ausgehändigt. Die Bedeutung der einzelnen Items und die entsprechenden Punktzahlen wurden erklärt. Die Studierenden erhielten zwei Beispielaufsätze und diskutierten die Bewertungen im Unterricht.

Eine weitere Quelle für Belege für die Validität sind verständliche und genaue Beschreibungen und Interpretationen der Bewertungen für die Studierenden [12]. In ihren Bewertungen gaben alle acht Studierenden an, dass die Aufgabe „vollständig“ oder „größtenteils verständlich“ war, und die Leitfragen wurden als „sehr hilfreich“ oder „hilfreich“ bewertet.

##### 4.2.3. Interne Struktur

Nach Fleiss kann eine Inter-Rater-Reliabilität (IRR) von 0,6 und mehr als ausgezeichnet und eine IRR über 0,4 und unter 0,6 als moderat (fair) bis gut angesehen werden. IRRs unter 0,4 werden als schlecht angesehen [15]. Dies bedeutet, dass die IRR für MaReS in unserer Studie nur für die Items 1 bis 4 und Item 9 akzeptabel sind. Betrachtet man den Wortlaut der Items, so scheint der Inhalt, der in den Items 1 bis 4 und 9 (allgemeine Verständlichkeit, Bezug zur Aufgabe, Situationsbeschreibung, Beschreibung der eigenen Emotionen, Auswahl externer Quellen) bewertet wird, konkreter und damit leichter zu bewerten zu sein als der der anderen Items. Die Items 5 bis 8 (Erklärung der eigenen Emotionen, Beschreibung der Perspektive des Gegenübers, Bezug der Perspektive des Gegenübers zur eigenen Perspektive, Einfluss früherer Erfahrungen und Überlegungen) und die Items 10 bis 12 (Bewertung der Situation, Handlungsstrategie, Erwartungen bezüglich der Anwendung der zukünftigen Handlungsstrategie/-en) sind integrale, aber schwierigere Bestandteile der Reflexion. Das Verständnis auf Seiten der Bewertenden scheint bei diesen Items subjektiver zu sein als bei den Items mit einer akzeptablen IRR. Einerseits kann die Erklärung der eigenen Emotionen schwierig sein und die Studierenden könnten das Gefühl haben, dass das Thema zu persönlich ist, um es in einem Reflexionsbericht zu teilen. Andererseits kann die Frage, ob eine Erklärung für Bewertende nachvollziehbar ist, von ihren persönlichen Erfahrungen abhängen. Das Gleiche gilt für den Perspektivwechsel. Die Items 10 bis 12 (Einschätzung der Situation, Handlungsplan) können für die Bewertenden schwierig sein, denn es kann sein, dass die Einschätzung der Situation durch die\*den Studierende\*n nachvollziehbar oder die Handlungsstrategie konkret ist,

der\*die Bewertende aber das Gefühl hat, dass sie nicht ausreicht, um eine ähnliche Situation in Zukunft besser zu bewältigen. Dies könnte einige Bewertende dazu veranlassen, niedrigere Bewertungen zu vergeben. Während der Bewertung der Aufsätze machten sich die Bewertenden Notizen zu Punkten, die schwierig zu bewerten waren. Die Analyse dieser Notizen ist nicht Teil dieser Studie, wird aber dazu beitragen, den Kodierleitfaden weiterzuentwickeln und wichtige Aspekte in zukünftigen Schulungen für Bewertende zu berücksichtigen.

Item 7 (die Perspektive des Gegenübers mit der eigenen Perspektive in Beziehung setzen) und Item 8 (Einfluss früherer Erfahrungen und Reflexionen) erhielten ebenfalls vergleichsweise niedrige Bewertungen. Dies könnte darauf hindeuten, dass diese Punkte zu schwierig sind. Im Falle von Item 7 könnten die Studierenden mehr Erklärungen, Beispiele und Training benötigen. Ein Grund für die niedrige Bewertung von Item 8 könnte sein, dass die Studierenden in unserer Studie nicht viel klinische Erfahrung hatten und es für sie schwierig war, auf frühere Erfahrungen oder Reflexionen zurückzugreifen.

Alle IRRs, die sich für das Instrument REFLECT ergeben haben, sind als schlecht zu bewerten. Während drei Studien hohe IRRs für REFLECT ergaben [39], [40], [57], war eine andere Studie ebenfalls nicht in der Lage, diese hohen IRRs zu replizieren. Die Autor\*innen der letztgenannten Studie geben an, dass der Unterschied in den Ergebnissen auf den Kontext zurückzuführen sein könnte, in dem die Belege für die Validität erhoben werden (z. B. andere Einrichtung und Studienpopulation) [18]. Da im Falle unserer Studie deutsche Muttersprachler\*innen ein englischsprachiges Instrument anwandten, sollten auch Sprachprobleme in Betracht gezogen werden. Es ist zudem wahrscheinlich, dass die Schulung der Bewertenden für REFLECT nicht ausreichend war. Es gibt zwar eine kurze Beschreibung der Anwendung von REFLECT in der Originalarbeit [57], aber die Bewertenden hatten keinen Zugang zu einem Kodierleitfaden. Unsere Schulung der Bewertenden für REFLECT bestand aus der Bewertung und Diskussion von nur einem Beispielaufsatz. Miller-Kuhlman et al. [39] wendeten für die Schulung Ihrer Bewertenden einen solideren Ansatz an: In einem Schulungszyklus verglichen die Bewertenden die Bewertungen und diskutierten Diskrepanzen für mehrere Beispielaufsätze und bewerteten dann weitere Beispiele, bis eine IRR von mindestens 0,8 erreicht war, bevor sie die Daten für ihre Studie sammelten. Diese Schulung für REFLECT erforderte sechs Stunden.

#### 4.2.4. Bezug zu anderen Variablen

Der Vergleich eines neu entwickelten Tests mit einem Test, von dem angenommen wird, dass er das gleiche Konstrukt misst, kann als Beleg zur Bestimmung der Konstruktvalidität herangezogen werden [1]. Die Autor\*innen fanden eine positive Korrelation für den Vergleich der MaReS-Bewertungen mit einer globalen Bewertung der Aufsätze und mit dem Instrument REFLECT, das zur Bewertung des Reflexionsniveaus der Studierenden ver-

wendet wird und ein individuelles schriftliches Feedback zur Förderung der Reflexionsfähigkeit geben soll [57]. Für die erste Runde von Aufsätzen zeigten sich starke und signifikante positive Korrelationen [8] für MaReS und REFLECT sowie für REFLECT und die globale Bewertung. Für die zweite Runde zeigten sich starke und signifikante positive Korrelationen für MaReS und das globale Rating sowie für REFLECT und die globale Bewertung. Nahezu moderate und moderat positive (nach Cohen als  $r$  von 0,3 [8] definiert) Korrelationen zeigten sich für MaReS/globale Bewertung im ersten Bericht und für MaReS/REFLECT im zweiten Bericht, die beide wahrscheinlich aufgrund der geringen Stichprobengröße nicht signifikant waren. Die Tatsache, dass in der ersten und zweiten Runde unterschiedliche Korrelationen gefunden wurden, könnte durch einen Wechsel der Bewertenden verursacht worden sein. Dennoch deuten die Ergebnisse darauf hin, dass MaReS dasselbe Konstrukt misst wie REFLECT [1], nämlich das Reflexionsniveau der Studierenden.

#### 4.2.5. Evaluationen

In dieser Studie werden Daten für Stufe 1 (Reaktion) des New-World-Kirkpatrick-Modells [26] untersucht. Bei der Betrachtung der Ergebnisse ist zu bedenken, dass die Anzahl der Studierenden, die die Evaluation für MaReS ausgefüllt haben, sehr gering war ( $n=8$ ) und Schlussfolgerungen mit Vorsicht gezogen werden müssen.

Einige Studierende teilen in ihren Reflexionsberichten sehr persönliche Geschichten und es muss darauf geachtet werden, dass das Feedback zu den Aufsätzen nicht wertend ist [21] und die Gefühle der Studierenden nicht verletzt. Der Inhalt der Reflexion ist subjektiv, weshalb Koole et al. vorschlagen, dass sich die Bewertung auf allgemeine Prozessfähigkeiten konzentrieren sollte [29]. Daher fokussieren die meisten Instrumente, die für die Bewertung von Reflexionsaufsätzen verwendet werden (einschließlich MaReS), auf den Prozess der Reflexion und nicht auf den Umgang mit einer Situation. Dennoch scheinen einige Studierende das Bedürfnis nach einem Kommentar zur Situation zu haben, denn eine Studierende erwähnte, dass das Feedback keine Auswirkungen auf ihn/sie hatte, weil nur die Struktur des Textes und nicht die Situation selbst bewertet wurde. Die Lehrenden sollten sich auch darüber im Klaren sein, dass die Aufsätze sehr persönliche Inhalte der Studierenden enthalten können. So könnten beispielsweise Beschreibungen von emotionalen Belastungen, Krankheiten oder traumatischen Erlebnissen in den Berichten enthalten sein. Daher sollten die Lehrenden im Voraus überlegen, wo sie Unterstützung finden können und wie sie in solchen Fällen reagieren werden.

Bei effektivem Feedback werden Informationen über frühere Leistungen genutzt, um eine positive und wünschenswerte Entwicklung zu fördern [2]. Die Studierenden empfanden das Feedback, das sie mit MaReS erhielten, als konstruktiv, hilfreich und nicht wertend.

Das Erreichen von Stufe 2 des New-World-Kirkpatrick-Modells [26] wird für MaReS eine Herausforderung dar-

stellen, da die Reflexionsfähigkeit kontextabhängig ist. Moniz et al. schließen aus ihrer Studie über die Verwendung von reflektierendem Schreiben zur Beurteilung von Schüler\*innen, dass für aussagekräftige Schlussfolgerungen über die Reflexionsfähigkeit etwa 14 Schreibproben pro Schüler\*in erforderlich sind, die jeweils von vier oder fünf Beurteilenden bewertet werden [40]. Diese Schlussfolgerung wirft Fragen über die Durchführbarkeit der summativen Bewertung von reflektierendem Schreiben auf. Auch die Ergebnisse unserer Studie weisen in die gleiche Richtung: Obwohl das Feedback-Instrument überwiegend positiv bewertet wurde, fühlte sich die Hälfte der Studierenden beim Verfassen ihres zweiten Aufsatzes etwas weniger sicher – und ihre Freitext-Antworten deuteten darauf hin, dass dieses Gefühl mit der Situation zusammenhing, die sie für die zweite Reflexion gewählt hatten (z. B. Probleme, ein Thema für die zweite Reflexion zu finden).

## 5. Schlussfolgerung

Die Magdeburger Reflexionsskala (MaReS) kann als Instrument verwendet werden, um das reflektierende Schreiben der Studierenden anzuleiten und strukturiertes Feedback in der Ausbildung von Gesundheitsberufen zu geben. In dieser Studie war die IRR für sieben der zwölf Items niedrig. Wir gehen davon aus, dass die Schulung der Bewertenden – bestehend aus der Bewertung und Diskussion eines exemplarischen Aufsatzes – nicht ausreichend war. Die Verwendung von mehr Aufsätzen für ein Training der Bewertenden und mehr Trainingszyklen würden wahrscheinlich zu höheren IRRs führen. Eine durchschnittliche Bewertungszeit von 13,9 Minuten scheint praktikabel und könnte sich verkürzen, wenn die Bewertenden mehr Erfahrung sammeln. Wenn Lehrende Reflexionsberichte und deren Bewertung in ein Curriculum einbinden möchten, muss bei der Planung der Ressourcen ausreichend Zeit für die Schulung der Bewertenden eingeplant werden. Vorsicht ist geboten, wenn eine summative Bewertung des reflektierenden Schreibens eingesetzt wird, unabhängig davon, welches Instrument verwendet wird, da die Vorhersagbarkeit des Ergebnisses von einem Aufsatz zum nächsten gering ist.

Die wenigen Studierenden, die ein Feedback zu MaReS gaben, hielten das Instrument für nachvollziehbar und hilfreich. Um diese Ergebnisse zu untermauern, sind weitere Studien mit einer größeren Anzahl von Studierenden erforderlich. Das Sammeln von Belegen für die Kirkpatrick-Stufen zwei und höher wird aufgrund der Kontextspezifität der Reflexionsfähigkeit eine Herausforderung darstellen.

Wir empfehlen MaReS als Instrument für den Unterricht und die formative Bewertung schriftlicher Reflexionen von Studierenden in der Ausbildung von Gesundheitsberufen, z. B. über klinische Erfahrungen während des Praktischen Jahres, der Famulatur oder der Blockausbildung, aber auch über Lernerfahrungen im Allgemeinen.

## ORCID der Autor\*innen

- Anja Härtl: [0009-0008-0818-6213]
- Stefan Rüttermann: 0000-0002-2293-8089
- Linn Hempel: [0009-0009-5421-2029]

## Danksagung

Wir bedanken uns bei den Mitgliedern des Ausschusses „Kommunikative und soziale Kompetenzen“ (KusK) und den Expert\*innen für die gute Zusammenarbeit und ihre Unterstützung.

## Interessenkonflikt

Die Autor\*innen erklären, dass sie keinen Interessenkonflikt im Zusammenhang mit diesem Artikel haben.

## Anhänge

Verfügbar unter <https://doi.org/10.3205/zma001752>

1. Anhang\_1.pdf (274 KB)  
Magdeburger Reflexionsskala

## Literatur

1. American Educational Research Association; American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999.
2. Archer JC. State of the science in health professional education: effective feedback. *Med Educ.* 2010;44(1):101-108. DOI: 10.1111/j.1365-2923.2009.03546.x
3. Aronson L, Niehaus B, Hill-Sakurai L, Lai C, O'Sullivan PS. A comparison of two methods of teaching reflective ability in Year 3 medical students. *Med Educ.* 2012;46(8):807-814. DOI: 10.1111/j.1365-2923.2012.04299.x
4. Aronson L, Niehaus B, Lindow J, Robertson PA, O'Sullivan PS. Development and pilot testing of a reflective learning guide for medical education. *Med Teach.* 2011;33(10):e515-e521. DOI: 10.3109/0142159X.2011.599894
5. Bok HG, Jaarsma DA, Teunissen PW, van der Vleuten CP, van Beukelen P. Development and validation of a competency framework for veterinarians. *J Vet Med Educ.* 2011;38(3):262-269. DOI: 10.3138/jvme.38.3.262
6. Bolton G. *Reflective Practice. Writing Professional Development.* Third Edition. London: SAGE; 2010.
7. Chu SY, Lin CW, Lin MJ, Wen CC. Psychosocial issues discovered through reflective group dialogue between medical students. *BMC Med Educ.* 2018;18(1):12. DOI: 10.1186/s12909-017-1114-x
8. Cohen J. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: L. Erlbaum Associates; 1988.
9. Cruess SR, Cruess RL, Steinert Y. Supporting the development of a professional identity: General principles. *Med Teach.* 2019;41(6):641-649. DOI: 10.1080/0142159X.2018.1536260

10. Devi V, Abraham RR, Kamath U. Teaching and Assessing Reflecting Skills among Undergraduate Medical Students Experiencing Research. *J Clin Diagn Res.* 2017;11(1):jc01-jc05. DOI: 10.7860/JCDR/2017/20186.9142
11. Devlin MJ, Mutnick A, Balmer D, Richards BF. Clerkship-based reflective writing: a rubric for feedback. *Med Educ.* 2010;44(11):1143-1144. DOI: 10.1111/j.1365-2923.2010.03815.x
12. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830–837. DOI: 10.1046/j.1365-2923.2003.01594.x
13. European Association of Establishments for Veterinary Education (EAEVE); Federation of Veterinarians of Europe (FVE). European System of Evaluation of Veterinary Training (ESEVT). Manual of Standard of Operation Procedure. Wien: EAEVE; 2019. Zugänglich unter/available from: [https://www.eaeve.org/fileadmin/downloads/SOP/ESEVT\\_SOP\\_2019\\_adopted\\_by\\_the\\_32nd\\_GA\\_in\\_Zagreb\\_on\\_30\\_May\\_2019.pdf](https://www.eaeve.org/fileadmin/downloads/SOP/ESEVT_SOP_2019_adopted_by_the_32nd_GA_in_Zagreb_on_30_May_2019.pdf)
14. Fink A. Survey Research Methods. In: Peterson P, Baker E, McGaw B, editors. *International Encyclopedia of Education*. Third Edition. Amsterdam: Elsevier Science; 2010. p.152-160. DOI: 10.1016/B978-0-08-044894-7.00296-7
15. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley-Interscience; 1981.
16. Frank J, Snell L, Sherbino J. *CanMEDS 2015 Physician Competency Framework*. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.
17. Fries JF. *Neue oder anthropologische Kritik der Vernunft*. Bd. 2. Zweite Auflage. 1831.
18. Grierson L, Winemaker S, Taniguchi A, Howard M, Marshall D, Zazulak J. The reliability characteristics of the REFLECT rubric for assessing reflective capacity through expressive writing assignments: A replication study. *Perspect Med Educ.* 2020;9(5):281-285. DOI: 10.1007/s40037-020-00611-2
19. Haramati A, Cotton S, Padmore JS, Wald HS, Weissinger PA. Strategies to promote resilience, empathy and well-being in the health professions: Insights from the 2015 CENTILE Conference. *Med Teach.* 2017;39(2):118-119. DOI: 10.1080/0142159X.2017.1279278
20. Heimes S. Warum Schreiben hilft: Die Wirksamkeitsnachweise zur Poesietherapie. Göttingen: Vandenhoeck&Ruprecht GmbH&Co.KG; 2012. DOI: 10.13109/9783666401619
21. Hewson MG, Little ML. Giving Feedback in Medical Education: verification of recommended techniques. *J Gen Intern Med.* 1998;13(2):111-116. DOI: 10.1046/j.1525-1497.1998.00027.x
22. Hunter DM, Jones RM, Randhawa BS. The use of holistic versus analytic scoring for large-scale assessment of writing. *Can J Program Eval.* 1996;11(2):61. DOI: 10.3138/cjpe.11.00
23. Johns C. Framing learning through reflection within Carper's fundamental ways of knowing in nursing. *J Adv Nurs.* 1995;22(2):226-234. DOI: 10.1046/j.1365-2648.1995.22020226.x
24. Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educ Res Rev.* 2007;2(2):130-144. DOI: 10.1016/j.edurev.2007.05.002
25. Kant I. *Kritik der reinen Vernunft*. 2. Auflage. Berlin: de Gruyter; 1968.
26. Kirkpatrick JD, Kirkpatrick WK. *Kirkpatrick's Four Levels of Training Evaluation*. Alexandria, VA: ATD Press; 2016.
27. Kiss A, Steiner C, Grossman P, Langewitz W, Tschudi P, Kiessling C. Students' satisfaction with general practitioners' feedback to their reflective writing: a randomized controlled trial. *Can Med Educ J.* 2017;8(4):e54-e59. DOI: 10.36834/cmej.36929
28. Koh YH, Wong ML, Lee JJ. Medical students' reflective writing about a task-based learning experience on public health communication. *Med Teach.* 2014;36(2):121-129. DOI: 10.3109/0142159X.2013.849329
29. Koole S, Dornan T, Aper L, Scherpbier A, Valcke M, Cohen-Schotanus J, Derese A. Factors confounding the assessment of reflection: a critical review. *BMC Med Educ.* 2011;11:104. DOI: 10.1186/1472-6920-11-104
30. Learman LA, Autry AM, O'Sullivan P. Reliability and validity of reflection exercises for obstetrics and gynecology residents. *Am J Obstet Gynecol.* 2008;198(4):461.e1-8. DOI: 10.1016/j.ajog.2007.12.021
31. Lim JY, Ong SY, Ng CY, Chan KL, Wu SY, So WZ, Tey GJ, Lam YX, Gao NL, Lim YX, Tay RY, Leong IT, Rahman ND, Chiam M, Lim C, Phua GL, Murugam V, Ong EK, Krishna LK. A systematic scoping review of reflective writing in medical education. *BMC Med Educ.* 2023;23(1):12. DOI: 10.1186/s12909-022-03924-4
32. Lovell BL, Lee RT. Burnout and health promotion in veterinary medicine. *Can Vet J.* 2013;54(8):790-791.
33. Lutz G, Scheffler C, Edelhaeuser F, Tauschel D, Neumann M. A reflective practice intervention for professional development, reduced stress and improved patient care – a qualitative developmental evaluation. *Patient Educ Couns.* 2013;92(3):337-345. DOI: 10.1016/j.pec.2013.03.020
34. Lyon P, Letschka P, Ainsworth T, Haq I. An exploratory study of the potential learning benefits for medical students in collaborative drawing: creativity, reflection and 'critical looking'. *BMC Med Educ.* 2013;13:86. DOI: 10.1186/1472-6920-13-86
35. Makoul G, Zick AB, Aakhus M, Neely KJ, Roemer PE. Using an online forum to encourage reflection about difficult conversations in medicine. *Patient Educ Couns.* 2010;79(1):83-86. DOI: 10.1016/j.pec.2009.07.027
36. *Medizinischer Fakultätentag (MFT). Nationaler Kompetenzbasierter Lernzielkatalog Medizin*. Berlin: MFT; 2015.
37. Messick S. Validity of test interpretation and use. In: Alkin MC, editor. *Encyclopedia of Educational Research*. 6th ed. New York: Macmillan; 1991.
38. Messick S. Validity. In: Linn RL, editor. *The American Council on Education/Macmillan series on higher education Educational measurement*. New York: Macmillan; 1989. p.13-103.
39. Miller-Kuhlman R, O'Sullivan PS, Aronson L. Essential steps in developing best practices to assess reflective skill: A comparison of two rubrics. *Med Teach.* 2016;38(1):75-81. DOI: 10.3109/0142159X.2015.1034662
40. Moniz T, Arntfield S, Miller K, Lingard L, Watling C, Regehr G. Considerations in the use of reflective writing for student assessment: issues of reliability and validity. *Med Educ.* 2015;49(9):901-908. DOI: 10.1111/medu.12771
41. Moon JA. *A handbook of reflective and experiential learning: Theory and practice*. London: Psychology Press; 2004.
42. Moskal BM. *Developing Classroom Performance Assessments and Scoring Rubrics-Part II*. ERIC Digest. 2003. Zugänglich unter/available from: <https://eric.ed.gov/?id=ED481715>
43. Moskal BM. Scoring rubrics: what, when and how? *Pract Assess Res Eval.* 2000;7(3). Zugänglich unter/available from: <http://pareonline.net/getvn.asp?v=7&n=34>
44. O'Sullivan P, Aronson L, Chittenden E, Niehaus B, Learman L. Reflective ability rubric and user guide. *MedEdPortal.* 2010. DOI: 10.15766/mep\_2374-8265.8133
45. OIE. OIE recommendations on the Competencies of graduating veterinarians ('Day 1 graduates') to assure National Veterinary Services of quality. Paris: OIE; 2012.



46. Ottenberg AL, Pasalic D, Bui GT, Pawlina W. An analysis of reflective writing early in the medical curriculum: The relationship between reflective capacity and academic achievement. *Med Teach.* 2016;38(7):724-729. DOI: 10.3109/0142159X.2015.1112890
47. Pee B, Woodman T, Fry H, Davenport ES. Appraising and assessing reflection in students' writing on a structured worksheet. *Med Educ.* 2002;36(6):575-585. DOI: 10.1046/j.1365-2923.2002.01227.x
48. Pennebaker JW. *Opening up: The healing power of expressing emotions.* London, New York: Guilford Press; 1997.
49. Randolph JJ. Free-marginal multirater kappa (multirater kfree): an alternative to Fleiss' fixed-Marginal multirater kappa. In: Paper presented at the Joensuu Learning and Instruction Symposium, Joensuu, Finland. 2005. Zugänglich unter/available from: [https://www.researchgate.net/publication/224890485\\_Free-Marginal\\_Multirater\\_Kappa\\_multirater\\_kfree\\_An\\_Alternative\\_to\\_Fleiss\\_Fixed-Marginal\\_Multirater\\_Kappa](https://www.researchgate.net/publication/224890485_Free-Marginal_Multirater_Kappa_multirater_kfree_An_Alternative_to_Fleiss_Fixed-Marginal_Multirater_Kappa)
50. Reis SP, Wald HS, Monroe AD, Borkan JM. Begin the BEGAN (The Brown Educational Guide to the Analysis of Narrative) – A framework for enhancing educational impact of faculty feedback to students' reflective writing. *Patient Educ Couns.* 2010;80(2):253-259. DOI: 10.1016/j.pec.2009.11.014
51. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. *Med Teach.* 2009;31(8):685-695. DOI: 10.1080/01421590903050374
52. Sandars J, Murray C. Digital storytelling for reflection in undergraduate medical education: a pilot study. *Educ Prim Care.* 2009;20(6):441-444. DOI: 10.1080/14739879.2009.11493832
53. Shapiro J, Rakhra P, Wong A. The stories they tell: How third year medical students portray patients, family members, physicians, and themselves in difficult encounters. *Med Teach.* 2016;38(10):1033-1040. DOI: 10.3109/0142159X.2016.1147535
54. Stevens DD, Cooper JE. *Journal Keeping: How to Use Reflective Writing for Learning, Teaching, Professional Insight and Positive Change.* Sterling (VA): Stylus Publishing; 2009.
55. Uygur J, Stuart E, De Paor M, Wallace E, Duffy S, O'Shea M, Smith S, Pawlikowska T. A Best Evidence in Medical Education systematic review to determine the most effective teaching methods that develop reflection in medical students: BEME Guide No. 51. *Med Teach.* 2019;41(1):3-16. DOI: 10.1080/0142159X.2018.1505037
56. Wald HS. Professional identity (trans)formation in medical education: reflection, relationship, resilience. *Acad Med.* 2015;90(6):701-706. DOI: 10.1097/ACM.0000000000000731
57. Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing. *Acad Med.* 2012;87(1):41-50. DOI: 10.1097/ACM.0b013e31823b55fa
58. Wald HS, Haramati A, Bachner YG, Urkin J. Promoting resiliency for interprofessional faculty and senior medical students: Outcomes of a workshop using mind-body medicine and interactive reflective writing. *Med Teach.* 2016;38(5):525-528. DOI: 10.3109/0142159X.2016.1150980
59. Wald HS, Reis SP, Borkan JM. Reflection rubric development: evaluating medical students' reflective writing. *Med Educ.* 2009;43(11):1110-1111.
60. West CP. Physician Well-Being: Expanding the Triple Aim. *J Gen Intern Med.* 2016;31(5):458-459. DOI: 10.1007/s11606-016-3641-2
61. Wong FK, Kember D, Chung LY, Yan L. Assessing the level of student reflection from reflective journals. *J Adv Nurs.* 1995;22(1):48-57. DOI: 10.1046/j.1365-2648.1995.22010048.x

**Korrespondenzadresse:**

Dr. Sabine Ramspott  
 Trillium GmbH Medizinischer Fachverlag, Jesenwanger  
 Str. 42b, 82284 Grafrath, Deutschland, Tel: +49  
 (0)176/76788212  
 sabine.ramspott@trillium.de

**Bitte zitieren als**

Ramspott S, Sonntag U, Härtl A, Rüttermann S, Roller D, Giesler M, Hempel L. *MaReS (Magdeburg Reflective Writing Scoring Rubric for Feedback) – development of a feedback method for reflective writing in health professions education: A pilot study in veterinary medicine.* *GMS J Med Educ.* 2025;42(2):Doc28.  
 DOI: 10.3205/zma001752, URN: urn:nbn:de:0183-zma001752

**Artikel online frei zugänglich unter**  
<https://doi.org/10.3205/zma001752>

**Eingereicht:** 12.10.2021  
**Überarbeitet:** 26.07.2024  
**Angenommen:** 16.12.2024  
**Veröffentlicht:** 15.04.2025

**Copyright**

©2025 Ramspott et al. Dieser Artikel ist ein Open-Access-Artikel und steht unter den Lizenzbedingungen der Creative Commons Attribution 4.0 License (Namensnennung). Lizenz-Angaben siehe <http://creativecommons.org/licenses/by/4.0/>.